DOCUMENT RESUME

ED 447 189 TM 032 077

AUTHOR	Lee, Won-Chan; Hanson, Bradley A.; Brennan, Robert L.
TITLE	Procedures for Computing Classification Consistency and Accuracy Indices with Multiple Categories. ACT Research Report Series.
INSTITUTION	American Coll. Testing Program, Iowa City, IA.
REPORT NO	ACT-RR-2000-10
PUB DATE	2000-10-00
NOTE	58p.
AVAILABLE FROM	ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243-0168.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Classification; *Estimation (Mathematics); *Item Response Theory; *Reliability
IDENTIFIERS	*Accuracy; Beta Binomial Test Model; Three Parameter Model

ABSTRACT

This paper describes procedures for estimating various indices of classification consistency and accuracy for multiple category classifications using data from a single test administration. The estimates of the classification consistency and accuracy indices are compared under three different psychometric models: the two-parameter beta binomial, four-parameter beta binomial, and three-parameter logistic item response theory (IRT) models. Using real data sets, the estimation procedures are illustrated, and the characteristics of the estimated indices are examined. This paper also examines the behavior of the estimated indices as a function of the latent variable. The IRT model tends to provide better fits to the data used in this study, and shows larger estimated consistency and accuracy. Although the results are not substantially different across different models, all three components of the models (i.e., the estimated true score distributions, fitted observed score distributions, and estimated conditional error variances) appear to have a great influence on estimates of the indices. Choosing a model in practice should be based on various considerations including the degree of the model fit to the data, suitability of the model assumptions, and the computational feasibility. (Contains 7 tables, 16 figures, and 37 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ACT Research Report Series 2000-10

ED 447 189

Procedures for Computing Classification Consistency and Accuracy Indices with Multiple Categories

Won-Chan Lee

Bradley A. Hanson

Robert L. Brennan

TM032077

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

ACT

October 2000

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

© 2000 by ACT, Inc. All rights reserved.

Procedures for Computing Classification Consistency and Accuracy Indices with Multiple Categories

Won-Chan Lee
Bradley A. Hanson
ACT, Inc.
Robert L. Brennan
The University of Iowa

Table of Contents

Page

Abstract iii

Acknowledgements iv

Introduction 1

Classification Consistency and Accuracy 3

Classification Consistency Indices 6

Classification Accuracy Indices 8

Models for Estimating Classification Indices 10

Beta Binomial Model 10

IRT Model 12

Illustrative Examples 13

Data 13

Model Fit 14

Marginal Classification Indices 15

Conditional Probabilities 18

Conclusions and Discussion 22

References 26

Tables 29

Figures 36

Abstract

This paper describes procedures for estimating various indices of classification consistency and accuracy for multiple category classifications using data from a single test administration. The estimates of the classification consistency and accuracy indices are compared under three different psychometric models: the two-parameter beta binomial, four-parameter beta binomial, and three-parameter logistic IRT models. Using real data sets, the estimation procedures are illustrated, and the characteristics of the estimated indices are examined. This paper also examines the behavior of the estimated indices as a function of the latent variable. The IRT model tends to provide better fits to the data used in this study, and shows larger estimated consistency and accuracy. Although the results are not substantially different across different models, all three components of the models (i.e., the estimated true score distributions, fitted observed score distributions, and estimated conditional error variances) appear to have a great influence on the estimates of the indices. Choosing a model in practice should be based on various considerations including the degree of the model fit to the data, suitability of the model assumptions, and the computational feasibility.

Acknowledgements

We are grateful to Matthew Schulz for discussions and many helpful comments and suggestions on an earlier version of the paper. The authors also wish to thank Debra Harris, Anne Fitzpatrick, Ann White, and Lin Wang for their helpful comments on the paper.

Procedures for Computing Classification Consistency and Accuracy Indices with Multiple Categories[1]

Introduction

It has been an important measurement practice in the context of mastery or competency testing to categorize examinees by mastery and non-mastery with respect to a specific standard. A number of studies have been devoted to quantifying reliability of mastery classifications (Huynh, 1976; Huynh & Saunders, 1980; Subkoviak, 1984; Berk, 1984). The term classification consistency is often referred to as reliability of classifications because the definition of classification consistency requires the concept of repeated testings, which constitutes the most essential component of reliability analyses (Feldt & Brennan, 1989). The importance of classification consistency has been widely recognized, and standard 2.15 in the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states that "When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure. . . ." Since data from such repeated testings are seldom available in practice, several writers proposed procedures for estimating classification consistency indices using test scores obtained from a single test administration by imposing psychometric models on test scores. Huynh (1976), Subkoviak (1984), and many others considered the beta binomial model. Huynh (1990), Wang, Kolen, and Harris (1996), and Schulz, Kolen, and Nicewander (1997, 1999) considered item response theory (IRT). Hanson and Brennan (1990) compared three different strong true-score models including the two-parameter beta binomial, the four-parameter beta binomial, and the four-parameter beta compound binomial.

---

[1] A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 2000.

The term classification accuracy is more closely related to validity of the classification system, evaluating the degree of accuracy of the classifications based on the observed scores as an attempt at classification based on true scores. Estimates of the various classification accuracy indices as well as classification consistency would be important for evaluating the psychometric properties of the classification process. Wilcox (1977) describes procedures for estimating the false positive and negative error rates under the beta binomial and beta compound binomial models. Huynh (1980) considers an asymptotic inferential procedure for the false positive and negative error rates. Brennan (1981) provides a summary of various statistical procedures for domain-referenced testing, and presents a step-by-step procedure for computing classification errors.

To date, most of the literature has been focused on binary classifications (i.e., mastery and non-mastery); there have been relatively few attempts to deal with the case in which each examinee is classified into one of three or more categories such as achievement levels based on a set of standards or cutoff scores. Some relevant references include Huynh (1978), who considered multiple classifications and reformulated Cohen's coefficient kappa. Schulz et al. (1997, 1999) presented a new procedure for defining multiple achievement levels, and evaluated the procedure in terms of classification consistency and classification errors. Livingston and Lewis (1995) presented a method for estimating classification consistency and accuracy for any type of test scores and provided some examples with multiple classifications. Wang et al. (1996) considered polytomous IRT models to estimate classification consistency indices for multiple categories defined based on polytomous items.

As an extension and generalization of the previous research (i.e., Hanson & Brennan 1990; Schulz et al., 1997, 1999), the present paper (1) describes procedures for estimating various indices of classification consistency and accuracy for multiple classifications from a single test administration; (2) compares the classification indices computed under three different psychometric models: the two-parameter beta binomial, four-parameter beta binomial, and IRT models; and (3) illustrates the procedures using real data sets. In addition, this paper examines

the behavior of the estimated classification indices as a function of the latent variable. These conditional probabilities will be very useful when reported in conjunction with the actual set of cutoff scores. The comparison of the IRT and two beta binomial models will show how the differences in the model assumptions would affect the results of the estimated classification indices. Note that Schulz et al. (1999) provide formulas for several classification indices under the IRT framework. In this paper, however, all the formulas and equations for the classification indices are presented in a more general way so that they can be applied to both the IRT and beta binomial models.

Classification Consistency and Accuracy

Suppose a testing procedure measures a single latent trait, ϕ, and let Φ denote a latent random variable. Further let $g(\phi)$ and Ω be the density and space of Φ. Assuming that the data to be modeled consist of test scores, x, from K dichotomously scored items, the marginal probability of the raw (i.e., number-correct) score is given by

$$\Pr(X = x) = \int_{\Omega} \Pr(X = x \mid \Phi = \phi) g(\phi) d\phi, \quad x = 0, 1, \ldots, K . \tag{1}$$

The marginal distribution $\Pr(X = x)$ is denoted here as $f(x)$, and the conditional error distribution, $\Pr(X = x \mid \Phi = \phi)$ is denoted as $f(x \mid \phi)$.

The general measurement situation considered here is one in which examinees are classified into one of H mutually exclusive categories on the basis of predetermined $H - 1$ observed score cutoffs, $c_1, c_2, \ldots, c_{H-1}$. Examinees with observed scores greater than or equal to zero, and less than c_1 will be classified into the first category, and so on. The Hth category consists of test scores between c_{H-1} and K. This method will be referred to here as observed category classifications. Now, let I_h $(h = 1, 2, \ldots, H)$ denote the hth category into which examinees with $c_{h-1} \le x < c_h$ are classified, and here $c_0 = 0$ and $c_H = K + 1$. Then, the conditional and marginal probabilities of each category classification are, respectively,

BEST COPY AVAILABLE

$$\Pr(X \in I_h \mid \Phi = \phi) = \sum_{x=c_{h-1}}^{c_h-1} f(x \mid \phi), \quad h = 1,2,...,H, \tag{2}$$

and

$$\Pr(X \in I_h) = \int_{\Omega} \sum_{x=c_{h-1}}^{c_h-1} f(x \mid \phi) g(\phi) d\phi, \quad h = 1,2,...,H. \tag{3}$$

Classification consistency is defined as the extent to which the classifications agree on the basis of two independent administrations of the test (or, two parallel forms of the test). Since such data from repeated measurements on a representative sample are rare, it has been customary to estimate classification consistency indices based on a psychometric model using test scores obtained from a single test administration. One principal output from the analysis of classification consistency with multiple categories is a symmetric marginal H X H contingency table. The elements of the H X H contingency table are composed of the joint probabilities of the row and column observed category classifications. Although it would be possible to derive a number of different indices from the multiple contingency table, two general indices of classification consistency are considered in this paper: the agreement index P and coefficient kappa (Cohen, 1960). The coefficient P is simply the sum of the diagonal elements in the H X H contingency table, and kappa, κ, is a consistency index adjusted for chance agreement. The probability of *in*consistent classification is $1 - P$. Further, suppose we separately apply each of $H - 1$ cutoff scores to the data as opposed to applying all the cutoffs at the same time. Then, we can obtain $H - 1$ sub-indices from analyses of $H - 1$ binary contingency tables such as P_m and κ_m where $m = 1,2,...,H - 1$. These sub-indices would be useful, for example, when one of the multiple cutoff scores is considered as a minimum competency level or a passing score. The classification consistency indices, P and P_m, can also be computed conditional on ϕ, which would provide some useful information for test users.

Classification accuracy refers to the extent to which the actual classifications using observed cutoff scores agree with the "true" classifications based on known true score cutoffs

(Livingston & Lewis, 1995). While classification consistency is defined on the basis of two observed score distributions on the two alternate forms of the test, classification accuracy is defined based on the bivariate distribution of the observed and true score distributions. To compute the classification accuracy indices, we need to specify true cutoff scores, $\phi_1, \phi_2, \ldots, \phi_{H-1}$, which segment the population into H categories based on examinees' true latent scores. We will refer to this method as true category classifications. Let Γ_l $(l = 1, 2, \ldots, H)$ denote true categories, which define the true status of examinees with $\phi_{l-1} \le \phi < \phi_l$. For $l = 1$, the condition is $\min(\phi) \le \phi < \phi_1$, and for $l = H$, $\phi_{H-1} \le \phi \le \max(\phi)$. As in the case of classification consistency, a marginal $H \times H$ contingency table can be produced, which contains the joint probabilities of observed and true category classifications, $\Pr(X \in I_h, \Phi \in \Gamma_l)$. The contingency table for classification accuracy is not symmetric, however, because the two distributions used to create the table are different. The overall classification accuracy index, symbolized here as γ, is defined as the sum of the diagonal elements in the $H \times H$ contingency table. The sum of the upper diagonal elements will indicate the overall probability of examinees' obtaining observed categories that are higher than their true categories, and will be denoted as P^+. Conversely, the sum of the lower diagonal elements will indicate the overall probability of examinees' obtaining observed categories that are lower than their true categories, and will be denoted as P^-. When each pair of the observed and true cutoff scores is applied separately, $H - 1$ sub-indices of γ, γ_m, can be computed, and P_m^+ and P_m^- will become what usually are called a false positive and false negative error rates.

Figure 1 presents diagrams for marginal classification consistency and accuracy indices when there are four cutoff scores. The 4x4 contingency tables are filled with the joint category probabilities. Each index is represented by the sum of the shaded cells of the corresponding contingency table. For example, the shaded cells in the tables in the first row correspond to various ways of counting consistent classifications. Note that the contingency tables for the accuracy indices are not symmetric, which leads to different values of the "+" and "-" error rates.

Classification Consistency Indices

We shall assume that, conditioned on ϕ, the two raw score random variables X_1 and X_2 on the two administrations of the test are independent and identically distributed. Then, the conditional joint distribution of X_1 and X_2 is given by

$$f(x_1, x_2 \mid \phi) = f(x_1 \mid \phi) f(x_2 \mid \phi). \tag{4}$$

The marginal joint distribution of X_1 and X_2 can be obtained by integrating the conditional probabilities in Equation 4 over the distribution of Φ:

$$f(x_1, x_2) = \int_{\Omega} f(x_1, x_2 \mid \phi) g(\phi) d\phi. \tag{5}$$

A consistent classification is made if both x_1 and x_2 for an examinee belong to the same category I_h. The conditional probability of falling in the same category on the two testing occasions is

$$\Pr(X_1 \in I_h, X_2 \in I_h \mid \Phi = \phi) = \left[\sum_{x_1 = c_{h-1}}^{c_h - 1} f(x_1 \mid \phi) \right]^2, \quad h = 1, 2, \ldots, H. \tag{6}$$

Then, the agreement index P conditional on ϕ is obtained by

$$P(\phi) = \sum_{h=1}^{H} \Pr(X_1 \in I_h, X_2 \in I_h \mid \Phi = \phi). \tag{7}$$

Applying each cutoff separately, sub-indices of P conditional on ϕ can be obtained as

$$P_m(\phi) = \left[\sum_{j=1}^{m} \Pr(X_1 \in I_j \mid \Phi = \phi) \right]^2 + \left[\sum_{j=m+1}^{H} \Pr(X_1 \in I_j \mid \Phi = \phi) \right]^2, \quad m = 1, 2, \ldots, H-1. \tag{8}$$

The marginal values of the agreement indices can be computed by

$$P = \int_{\Omega} P(\phi) g(\phi) d(\phi), \tag{9}$$

and

$$P_m = \int_{\Omega} P_m(\phi) g(\phi) d(\phi), \quad m = 1, 2, \ldots, H-1. \tag{10}$$

The coefficients P and P_m represent the probability that a randomly selected examinee is classified in the same observed category on the two testing occasions. The probabilities of inconsistent classifications can be obtained by subtracting the probabilities of consistent classifications from one.

The overall coefficient kappa when applying all cutoff scores together is

$$\kappa = \frac{P - P_c}{1 - P_c}, \tag{11}$$

where P_c is the probability of consistent classification by chance. The chance agreement is the sum of squared marginal probabilities of each category classification, which is written as

$$P_c = \sum_{h=1}^{H} \Pr(X_1 \in I_h) \Pr(X_2 \in I_h) = \sum_{h=1}^{H} [\Pr(X_1 \in I_h)]^2 . \tag{12}$$

The probability P_c is determined under two complete random assignment procedures, in each of which examinees are assigned to a category according to the rule of the marginal category probabilities. Note that since $1/H \le P_c \le P$, κ could be viewed as a rescaled-version of P such that $0 \le \kappa \le 1$. (See Huynh, 1978 for more detailed properties of κ.) It may also be noted that P_c and κ are not applicable to a particular individual examinee. Agreement by chance conditional on ϕ does not make sense either conceptually or mathematically--the sum of the squared probabilities of each category classification given ϕ will be identical to $P(\phi)$. Finally, applying each cutoff separately, sub-indices of κ are obtained as

$$\kappa_m = \frac{P_m - P_{mc}}{1 - P_{mc}}, \tag{13}$$

where

$$P_{mc} = \left[\sum_{j=1}^{m} \Pr(X_1 \in I_j)\right]^2 + \left[\sum_{j=m+1}^{H} \Pr(X_1 \in I_j)\right]^2, \quad m = 1, 2, ..., H-1. \tag{14}$$

Classification Accuracy Indices

Suppose an examinee has an observed and a latent score: $x \in I_h$ $(h = 1, 2, ..., H)$ and $\phi \in \Gamma_l$ $(l = 1, 2, ..., H)$. An accurate classification is made when $h = l$. The conditional probability of accurate classifications is given by

$$\gamma(\phi) = \Pr(X \in I_l \mid \Phi = \phi), \tag{15}$$

where $l (= 1, 2, ..., H)$ is the category such that $\phi \in \Gamma_l$, and $\Pr(X \in I_l \mid \Phi = \phi)$ is computed using Equation 2.

Let $P^+(\phi)$ refer to the conditional probability that an examinee with a latent score ϕ within the range of a true category obtains an observed score falling in an observed category, which is one or more higher than the true category. Likewise, $P^-(\phi)$ refers to the conditional probability of an examinee's getting an observed category one or more lower than the true category. These two conditional error rates, respectively, are

$$P^+(\phi) = \sum_{h=l+1}^{H} \Pr(X \in I_h \mid \Phi = \phi), \tag{16}$$

and

$$P^-(\phi) = \sum_{h=1}^{l-1} \Pr(X \in I_h \mid \Phi = \phi), \tag{17}$$

where $l(=1,2,...,H)$ is the category such that $\phi \in \Gamma_l$. Note that the conditional accuracy indices in Equations 15, 16, and 17 are based on the single conditional distribution of the observed number-correct scores. The marginal classification accuracy index, γ, and the marginal error rates, P^+ and P^- are obtained by integrating the corresponding conditional indices over the distribution of Φ. The index γ refers to the probability that a randomly chosen examinee with a latent score falling in a true category is classified in the same observed category as the true category. The marginal error rates P^+ and P^- can be interpreted in a similar manner as the conditional counterparts except that now the probability is for a randomly chosen examinee.

As in the previous section for the classification consistency indices, we can also compute sub-indices of γ when applying each cutoff separately as:

$$\gamma_m(\phi) = \begin{cases} \sum_{h=1}^{m} \Pr(X \in I_h \mid \Phi = \phi), & \phi \in \Gamma_l \text{ and } l \le m \\ \sum_{h=m+1}^{H} \Pr(X \in I_h \mid \Phi = \phi), & \phi \in \Gamma_l \text{ and } l > m \end{cases} \tag{18}$$

for $m = 1,2,...,H-1$. Similarly, sub-indices of the conditional error rates are obtained by

$$P_m^+(\phi) = \sum_{h=m+1}^{H} \Pr(X \in I_h \mid \Phi = \phi), \quad \phi \in \Gamma_l \text{ and } l \le m, \tag{19}$$

and

$$P_m^-(\phi) = \sum_{h=1}^{m} \Pr(X \in I_h \mid \Phi = \phi), \quad \phi \in \Gamma_l \text{ and } l > m \tag{20}$$

for $m = 1,2,...,H-1$. The marginal probabilities of γ_m, P_m^+, and P_m^- are obtained by integrating Equations 18, 19, and 20 over Φ distribution, and the resulting P_m^+ and P_m^- are comparable with what people traditionally call a false positive and false negative error rates.

Models for Estimating Classification Indices

Three different models are considered in this paper for estimating the classification indices described in the previous sections: the two-parameter beta binomial (2PB), four-parameter beta binomial (4PB), and three-parameter logistic IRT models. The assumptions involved in the three models (especially between the IRT and two binomial models) are quite different, and have important consequences in the interpretation of the results in this study.

In general, the 4PB model is known to outperform the 2PB model in fitting distributions of observed test scores. Two reasons for including the 2PB model in this study are: (a) the 2PB model is simple and often provides adequate fits in many cases, and (b) fitting the 4PB model sometimes provides an upper limit of the estimated true score distribution that is less than one or more of the higher true score cutoffs, which would cause some zero probabilities in contingency tables. The 2PB model is free from this problem because the 2PB model always sets the lower and upper limit of the beta distribution at 0 and 1, respectively.

The basic role of the models in estimating classification indices is to estimate the latent score distribution and predict the observed score distribution. Once the latent and observed score distributions are estimated, $H \times H$ contingency tables can be created, which, in turn, are used as a basis for computing the classification indices considered in this paper. Note that the parameters of the models, distributions of latent and observed scores, and all the classification indices are estimated based on actual data from a single test administration.

Beta Binomial Model

For the beta binomial model, the latent score ϕ in all preceding equations is replaced with the true proportion-correct score τ. Under the 2PB model, the conditional distribution of X given τ is assumed to be binomially distributed, and the density of τ is assumed to be two-parameter beta with two shape parameters α and β (Keats & Lord, 1962). The parameter space of τ for the 2PB model is $0 \leq \tau \leq 1$. The parameters α and β can be estimated using KR21 and the first two moments of the actual observed score distribution. Then, given the two

parameters of the τ distribution and the assumption of the binomially distributed errors, the observed score distribution can be estimated. Hanson (1991) provides a closed-form formula to compute the observed score distribution, which can be used for both the 2PB and 4PB models.

The 4PB model considers the beta true score distribution with four parameters: two shape parameters (α and β) and the lower and upper limits of the distribution (a and b), which define the parameter space for τ, $a \leq \tau \leq b$. As with the 2PB model, the conditional error distribution is assumed to be binomial. Although a more complicated model than the binomial can be used for the error distribution (two-term approximation to the compound binomial), Hanson and Brennan (1990) found very little difference between results using binomial or the two-term approximation to the compound binomial as error distributions. Using the actual observed test scores, the true score distribution can be estimated by the method of moments (Lord, 1965). First, the first four moments of the true score distribution are estimated from the first four moments of the actual observed score distribution. Then, the four parameters (α, β, a, and b) of the true score distribution are estimated using the expressions of the parameters in terms of the first four moments. The fitted observed score distribution is obtained based on the estimates of the four beta parameters using the closed-form formula in Hanson (1991). More detailed explanations about estimating the beta parameters, the observed score distribution, and classification consistency indices for binary cases are provided in Hanson (1991).

To evaluate the behavior of some of the indices across the true score distribution, it is informative to report some conditional probabilities or distributions of interest such as conditional probabilities of inconsistent classifications, accurate classifications, error rates, and observed categories. Moreover, as discussed later, some conditional probabilities are useful to assess validity of the classification system with a particular set of cutoff scores. To estimate conditional distributions, it is convenient to use a set of several discrete true score points. For example, with a set of discrete values of τ, each τ value can be substituted into the closed-form formula in Hanson (1991) to obtain the conditional observed score distribution given τ. Other

relevant conditional probabilities can be obtained in a similar manner. The method used in this paper to find a set of discrete true scores is described in next section.

IRT Model

Although the basic assumptions of the beta binomial and IRT models are quite different, the general framework and formulas for the classification consistency and accuracy indices hold for both models. For the IRT model, the ability parameter θ and the latent random variable Θ are used in place of ϕ and Φ in all previous formulas for the classification indices. The conditional probability of X given θ, $f(x|\theta)$, is a function of item parameters, and can be represented by a compound binomial distribution (Lord, 1980). These conditional observed score distributions can be computed using a recursive algorithm in Lord and Wingersky (1984). Kolen and Brennan (1995, pp. 182-183) provides an illustrative example for using the recursive formula.

The integrals in equations (e.g., Equations 1, 3, 5, 9, and 10) can be evaluated by quadrature for the density of θ instead of using the whole space of θ, $-\infty < \theta < \infty$ (Press, Teukolsky, Vetterling, & Flannery, 1992). If a discrete distribution of θ is used, the integral in Equation 1 for the marginal observed score distribution becomes a sum:

$$\Pr(X = x) = \sum_{r=1}^{R} \Pr(X = x | \Theta = q_r)\pi_r, \quad x = 0,1, ..., K, \tag{21}$$

where there are R discrete values of θ given by $q_1, q_2, ..., q_r$, and $\pi_r = \Pr(\Theta = q_r)$. All the other equations containing an integral can be replaced with summations in the same manner. In this paper, the quadrature points and posterior weights for the θ distribution that are output in Phase 2 of BILOG3 (Mislevy & Bock, 1990) are used for the values of q_r and π_r, respectively, and 40 quadrature points are employed.

To achieve comparability between the beta binomial model and the IRT model, the θ metric can be transformed to the τ metric through the test characteristic curve:

$$\tau = \sum_{i=1}^{K} \Pr(U_i = 1 \mid \Theta = \theta), \tag{22}$$

where U_i is the random variable representing the response to item i. The 40 θ values are transformed, using Equation 22, to τ values that are used for the beta binomial models. All the results in this paper pertaining to the conditional probabilities are reported on the metric of τ.

Illustrative Examples

The procedures for estimating classification consistency and accuracy indices are illustrated using real data sets, and results are compared for the three different models. As Hanson and Brennan (1990) recommended, the fit of the three models to each data set will be evaluated first. Results reported here include some plots of conditional probabilities as well as tabulated marginal classification indices.

Data

The Work Keys assessments developed by ACT, Inc. include eight tests designed to measure eight different areas of job skills (ACT, Inc., 1998). Data used for the examples in this paper are from four forms of *Locating Information* and three forms of *Applied Mathematics* administered for equating in Fall, 1997. The sample sizes are about 3,000 ranging from 2,918 to 3,275 except for one form of *Applied Mathematics*, which was administered to 19,158 examinees. Each form of *Locating Information* contains 32 dichotomous items, and each form of *Applied Mathematics* contains 30 dichotomous items. The number of categories for the two tests is five for *Locating Information* and six for *Applied Mathematics*.

For the Work Keys assessments, the skill levels are defined via a number-correct scoring procedure, and the θ cutoffs are determined based on the IRT-estimated domain scores using all items in the pool (Schulz et al., 1999). The observed cutoff scores for each form are determined by searching for the integer number-correct scores that are as close to the θ cutoffs as possible

through the test characteristic curve relationship. Since the scale of the initial set of θ cutoffs is different from the scale of the item parameters for the data used here, the θ cutoffs need to be rescaled for each test. The Stocking and Lord (1983) scale transformation procedure is employed to obtain the transformation parameters. The rescaled θ cutoffs are then transformed to true scores to be used for the beta binomial models. Table 1 shows the final sets of θ cutoffs and form-specific true and observed score cutoffs for the two tests. The true score and observed (proportion-correct) score cutoffs are very close in general, with a maximum difference of .04. The discrete distribution of θ is obtained from BILOG3 Phase 2 output based on one form of each test.

Note that these data are used for convenience and illustrative purposes only, and the results reported here should not be viewed from any other standpoint. The purpose of the present study is *not* to evaluate the current classification system of the particular program, but to describe the procedures to compute various classification indices, discuss how to analyze the results, and compare results from three different underlying models.

Model Fit

The actual observed score distributions and fitted observed score distributions for the three models are plotted in Figures 2 and 3 for the two tests. In general, the IRT model appears to provide fitted distributions that are closest to the actual observed score distributions. The fitted distributions of the 4PB model are very close to the actual observed score distributions in many cases, but sometimes show bias (e.g., Form A of *Locating Information* in Figure 2). The 2PB model exhibits relatively inferior fits to the data compared to the other two models. However, the fit of the 2PB model is very similar to that of the 4PB model in such cases as Form D of *Locating Information* and Form X of *Applied Mathematics*. Notice that the fitted observed score distributions for the three models are relatively close to each other for Forms D, X, and Z. In general, all three models appear to provide better fits to the data for *Applied Mathematics* than

for *Locating Information*. It may be expected that the similarity in fitted distributions would result in similar estimated classification indices.

To explore further the fit of the models, the actual observed proportions of examinees and the estimated proportions of examinees for each category are computed and displayed in Tables 2 and 3 for the two tests. The boldfaced numbers indicate estimated category proportions that are closest to the actual values across models, and the underlined numbers indicate estimated proportions that are most different from the actual ones. As noted previously, the estimated proportions under the IRT model are generally closer to the actual ones than those under the other two models. The 2PB model, in general, shows the worst fitted proportions.

The better fit of the 4PB model than the 2PB model is consistent with findings in previous research (Hanson & Brennan, 1990). The better fit of the IRT model relative to the fit of the other models might be due to the fact that the IRT model fits item scores (i.e., more parameters) as opposed to test scores--the beta binomial models fit test scores to the data. In addition, scoring and scaling for the Work Keys assessments are based on IRT (ACT, Inc., 1999). The degree of the fit of a model to data is crucial in estimating classification consistency and accuracy indices, and it will affect the actual values for the estimates of the indices as shown later.

Marginal Classification Indices

Table 4 presents the marginal classification consistency indices, P, κ, and P_c with five different classification types for the four forms of *Locating Information*. The classification type where all cutoffs are applied at the same time is labeled "All", and $m = 1,...,4$ indicates dichotomous situations applying only one cutoff at a time. The numbers have been rounded, and thus, for example, 1.0 and 0.0 should be read "very close to one" and "very close to zero."

In general, the estimates of P do not greatly differ for the three models--the maximum difference is .09 between the IRT and 2PB models for Form C. The difference between κ estimates is more substantial. Except for a few cases, the estimates of P and κ are largest for the

IRT model. Even if the IRT model produces the largest P and κ estimates, it is the 2PB model that shows the smallest values of the P_c estimates in almost all cases. The estimates of P when all cutoffs are applied at the same time are smaller than any estimates of P_m, which is consistent with the conjecture that it would be harder to make consistent classifications with more categories. Moreover, P_m tends to be relatively small for mid values of m and large for low and high values of m. This is also understandable because the conditional measurement errors for number-correct scores are supposed to be larger near the middle of the score distribution than near both extremes. As m increases, P_c and P tend to change in a similar pattern and are positively correlated. However, as a function of both P_c and P, κ tends to decrease as m increases. Lastly, notice that the estimates of κ for the 4PB model when $m = 3$ and 4 are exceptionally small, which is related to the shape of the estimated true score distribution for the 4PB model. This issue is discussed next in more detail in the presentation of results for the classification accuracy indices.

Table 5 summarizes the estimated marginal classification accuracy indices for *Locating Information*. The trend of the estimated γ appears to be similar to that of estimated P. The estimates of γ are smaller when all cutoffs are applied together than those for any dichotomous cases. Also, the estimates of γ_m are smallest when $m = 2$. Among the three models, the IRT model yields the highest γ for most cases. In other words, the sum of P^+ and P^- are smallest for the IRT model. The 4PB model tends to produce the estimates of γ that are smaller than those for the IRT model, but larger than those for the 2PB model. It can also be observed that the percentage of P^- that accounts for the sum of the two error rates, P^+ and P^-, is highest for the 4PB model, which appears to be due to different shapes of the estimated true score distributions for the 4PB model compared to the other two models.

Even though the shapes (not the fit) of the fitted observed score distributions for the three models (i.e., bell-shaped) are similar, the shapes of the estimated true score distributions for the 4PB model are very different from those for the other two models. Figure 4 presents plots of estimated true score distributions for the four forms of *Locating Information*. The estimated true

score distribution for the IRT model is obtained by transforming the 40 discrete θ quadrature points to true scores using Equation 22 and directly using the posterior weights for θ as density. The posterior distribution of θ for the IRT model is obtained by using the default BILOG3 options, which employs the standard normal distribution as a starting point. Note that the true score distribution for the IRT model is provided only for one form because the true score distributions for the other forms are nearly indistinguishable. While the shapes of the estimated true score distributions for the IRT and 2PB models are close to the normal distribution, those for the 4PB model are severely negatively skewed in many cases, and sometimes show a J-shape. The extremely negatively-skewed true score distributions for the 4PB model are the result of one of the shape parameters being less than one.

In addition, the upper limits of the estimated true score distributions for the 4PB model, in many cases, turn out to be less than one and sometimes even smaller than some of the higher true score cutoffs. A severely negatively-skewed true score distribution with a very low upper limit matched with a bell-shaped observed score distribution would necessarily produce high rates of P^-. In Table 5, the values of P^- for the 4PB model when $m = 3$ and 4 are truly zero before rounding, because the true score cutoffs for $m = 3$ and 4 exceed the upper limits of the true score distributions. This is true only when $m = 4$ for Form D, which has an estimated true score distribution that is much less negatively skewed than the other forms. The zero density for true scores above the upper limit of the distribution may not be a critical issue because the density of such high true scores will be very close to zero anyway as seen in the results of the other two models. Note that the foregoing discussion about the upper limit parameter does not apply to the 2PB model, which sets the lower and upper limits of the true score distribution at zero and one.

Tables 6 and 7 contain results analogous to those presented in Tables 4 and 5 for *Applied Mathematics*. Most observations made in Tables 4 and 5 are still valid for Tables 6 and 7 with a few exceptions. In Tables 6 and 7, the estimates of the classification consistency and accuracy indices for the 4PB and 2PB models are, in general, very similar to each other, in contrast to the

differences seen in Tables 4 and 5. This appears to be a consequence of the similar fits of the two models to *Applied Mathematics* data as shown in Figure 3. Moreover, Figure 5 shows that, for *Applied Mathematics*, the estimated true score distributions for the 4PB model are more similar to those for the 2PB model showing less negatively skewed shapes than they are for *Locating Information*.

Conditional Probabilities

The results concerning conditional probabilities are presented here graphically. Although the results are displayed conditional on true scores, the same information could be obtained conditional on θ. Note that since the 2PB and 4PB models use the same model for errors, which is binomial, the conditional observed score distribution given a true score is the same for the two models. The same conditional observed score distribution will, in turn, lead to the same conditional classification consistency and accuracy indices. Thus, a general term "beta binomial model" is used to refer to the 2PB and 4PB models for the subsequent results.

Figures 6 and 7 contain plots of estimated conditional probabilities of inconsistent classifications [i.e., $1 - P(\tau)$ in Equation 7] for different forms of the two tests. The number of humps is consistent with the number of cutoffs for each test, in general, with the peak of each hump corresponding to each observed proportion-correct cutoff score, which indicates that inconsistent classifications are more likely to occur for the examinees with true scores near the observed cutoffs. On the other hand, the probability of inconsistent classifications tends to be minimal in the middle of the true score distribution for each observed category. The conditional probability of inconsistent classifications decreases as the true score falls farther from the observed cutoffs. Also notice that the beta binomial model provides probabilities of inconsistent classifications that are always larger than the IRT model, and the differences are more notable at true score levels near the local minima. As discussed later, these differences are related to the difference between the assumptions of the two models.

Figures 8 through 11 present plots of estimated conditional probabilities of falling within each observed category (i.e., Equation 2) using the beta binomial and IRT models for the two tests. The five (six for *Applied Mathematics*) solid lines in Figures 8 and 9 represent the conditional probabilities of the observed categories. As anticipated, the peak of each category probability falls within the range of the corresponding true category. On the whole, the results for the IRT model are very similar to those for the beta binomial model. However, the curves for the IRT model are somewhat higher indicating higher estimates of accurate classifications in general.

Each plot in Figures 8 – 11 can be interpreted in the following way. If a vertical line is drawn at a true score, the line will meet the curve for the true category corresponding to that score as well as some of the other curves. (Strictly speaking, the line meets all the curves, but some of the curves have zero probabilities.) The height of the point where the vertical line crosses the curve corresponding to the true category will be the conditional probability of accurate classifications, $\gamma(\tau)$ (Equation 15). The crossing point of the line and any curve lower than the true category can be found, and all those points will add up to the probability of classifications lower than the true level, $P^{-}(\tau)$ (Equation 17). Likewise, the sum of crossing points of the line and the curves above the true category will be equal to $P^{+}(\tau)$ (Equation 16). For example, the estimates of $\gamma(\tau)$, $P^{+}(\tau)$, and $P^{-}(\tau)$ for an examinee with $\tau = .6$ within a true level of 2 on Form A of *Locating Information* (Figure 8) are about .70, .25, and .05, respectively.

As another example, suppose an examinee has a true score of .84 for Form A of *Locating Information* in Figure 8. The true score of .84 is in the true Level 3 and also located in between the third true cutoff and the crossing point of the third and fourth observed category curves. For this particular examinee, the probability of obtaining the fourth observed category is higher than obtaining the third one (i.e., accurate classification). Since most of the true-score cutoffs are positioned slightly to the right of the crossing points, examinees with true scores falling in the

small areas between the true cutoffs and the crossing points will have higher $P^{+}(\tau)$ than $\gamma(\tau)$, which is not preferable.

Although it would be very difficult to achieve, a statistically "preferable" set of true score cutoffs corresponding to a set of observed score cutoffs would be set at the crossing points of the curves. Then, at least four desirable properties can be achieved assuming that the conditional probabilities of falling within each observed category are symmetric: (1) the probabilities of accurate classifications and one-less-than-true-level classifications will be the same at any true score cutoff, (2) the peak of any observed category probability will be located approximately in the middle of the true score interval for the particular category, (3) the probability of accurate classifications will be higher than any error rates across the entire distribution of true scores, and (4) the overall positive and negative error rates will be approximately the same. Note that the focus of the discussion here is only on the statistical properties of the cutoff scores. Indeed, most of the commonly used standard-setting methods require information about test content and examinees' performance (Berk, 1996). Reviews of the numerous standard-setting methods are presented in Berk (1986, 1996), Shepard (1980, 1984), and Kane (1994). Recently, Reckase (2000) provides a summary of the process used to set the standards on the National Assessment of Educational Progress (NAEP).

For a better graphical view, the three conditional classification accuracy indices, $\gamma(\tau)$, $P^{-}(\tau)$, and $P^{+}(\tau)$, are plotted in Figures 12 through 15 for the two models and two tests. Notice that there are several ranges of true scores where $P^{+}(\tau)$ is larger than $\gamma(\tau)$, which exactly correspond to the small areas discussed in the previous paragraph. In all cases, $P^{+}(\tau)$ and $P^{-}(\tau)$ exhibit a discernible trend. That is, $P^{+}(\tau)$ tends to rise as the true score approaches the next true score cutoff; and at each true score cutoff, $P^{+}(\tau)$ suddenly drops down to be a minimum and $P^{-}(\tau)$ becomes a maximum. As the true score diverges from each true score cutoff, $P^{-}(\tau)$ decreases while $P^{+}(\tau)$ increases until the next true score cutoff.

Figures 8 - 11 and 12 – 15, taken as a whole, seem to suggest that $P^{+}(\tau)$ is larger than $P^{-}(\tau)$, and the true cutoffs are somewhat higher than the corresponding observed score cutoffs.

It should be noted, however, that the mismatch between the true cutoffs and the crossing points, shown in Figures 8 - 11, still exists even when the true cutoff and the corresponding observed cutoff are exactly the same. For example, the true and observed score cutoffs of Level 4 for Form Y, *Applied Mathematics* are the same (see Table 1), but the true score cutoff still does not match with the crossing point (see Figures 10 and 11). The less preferable sets of cutoffs are not merely due to the use of integer observed cutoffs as an approximation to the non-integer true score cutoffs. As already seen, even exactly the same true and observed score cutoffs do not necessarily constitute preferable (as defined in this paper) sets of cutoff scores.

Figure 16 illustrates, in a different way, the discrepancy between the positive and negative error rates given the true and observed score cutoffs for Form Y of *Applied Mathematics*. Some selected conditional probabilities of observed scores near the true score cutoffs for Levels 2, 4, and 6 are plotted. The dotted lines are associated with the observed score cutoffs--the observed score cutoffs for Levels 2, 4, and 6 are 12, 21, and 28, respectively. Notice that the true score cutoffs for Levels 2 and 4 almost exactly correspond to the peaks of the conditional probability curves associated with the observed score cutoffs. Since the observed and true score cutoffs for Levels 2 and 4 are almost equal (see Table 1), the probabilities of the observed scores corresponding to the observed score cutoffs are expected to be maximum at the true score cutoff points. Assuming symmetric shapes of the observed score probabilities, which is in fact the case in this example, this approach necessarily produces differences between the positive and negative error rates. For instance, focusing on Level 2, examinees with true scores under the shaded area labeled as "A" have positive error rates (associated with the observed score of 12) represented by the height of the $\Pr(X = 12)$ curve because the height is the probability of obtaining the observed score of 12 (i.e., observed level of 2) for true scores lower than the true cutoff (i.e., true level of 1). By contrast, examinees with true scores under the area labeled "B" show negative error rates (associated with the observed score of 11) represented by the height of the $\Pr(X = 11)$ curve, which is the probability of obtaining the observed score of 11 (i.e., observed level of 1) for true scores higher than the true cutoff (i.e., true level of 2). (Note that we consider only two observed

scores near the true cutoff to make things simpler, but adding all the other probability curves would lead to the same conclusion.) The overall magnitude of the positive error rates is clearly greater than that of the negative error rates. The two error rates would be approximately equal in general, if the true score cutoff for Level h were in the middle of the peaks of the two conditional probability curves associated with the observed score cutoff c_h and $c_h - 1$. Unlike Level 2 and 4, the true score cutoff for Level 6 is placed in the left of the middle of the two curves, which leads to larger negative error rates for examinees with high true scores (see also Figure 14).

Conclusions and Discussion

Using the IRT and two beta binomial models as a psychometric tool, this paper presents formulas for the various indices of classification consistency and accuracy for multiple classifications based on test scores obtained from a single test administration. The results of this study indicate that all three components of the models (i.e., the estimated true score distributions, fitted observed score distributions, and estimated conditional error variances) had a great influence on the estimates of the classification indices. From the examples presented in this paper, it was found that the IRT model provided a better fit to the data than the 4PB model, which, in turn, provided a better fit than the 2PB model. Consistent with findings from the previous study (Hanson and Brennan, 1990), the 2PB model tended to produce inadequate fits to the data in some cases. Although the results were not substantially different across different models, the IRT model appeared to produce somewhat higher estimates of classification consistency and accuracy.

The marginal and conditional probabilities of inconsistent classifications for the beta binomial model were larger than those for the IRT model in most cases. Some plausible explanations for the phenomenon are: (1) the conditional error variance for the beta binomial model is larger than that of the IRT model, (2) determination of the achievement levels for the Work Keys assessments is based on IRT, and (3) the IRT model provides better fits to the data. The larger conditional error variance of the beta binomial model is primarily due to the

differences in the assumptions of the two models as discussed in Lee, Brennan, and Kolen (2000) and Lord (1984)--the beta binomial model assumes randomly parallel forms of a test allowing an additional source of errors due to form variation, as opposed to the IRT assumption of strictly parallel forms, which involves a conceptual replication of a test with a set of items having identical item parameters. It seems reasonable to presume that the larger conditional error variance is associated with the larger classification errors.

It was also found that the difference between the two models in probabilities of inconsistent classifications (Figures 6 and 7) was greater near the middle of the true score distribution for each category, where the classifications were most consistent. This is to be expected in that the magnitude of the conditional error variance (e.g., the width of the observed score distribution) for an individual with a true score near a cutoff score does not have much influence on the probability of classifications. By contrast, the magnitude of the conditional error variance more substantially affects the probability of consistent classifications for an individual with a true score near the middle of a category. For example, longer tails of the observed score distribution located in the middle of a category would cause more classification errors. The same sort of argument applies to the case of classification accuracy.

The 4PB model yielded estimates of classification indices that were different from those for the 2PB and IRT models in terms of (1) the small estimated classification consistency for dichotomous classifications with high cutoffs, and (2) the large percentage of the negative error rates that accounts for the total error rates. The primary reason was that the estimated true score distributions for the 4PB model were quite different from those for the IRT and 2PB models, while the fitted observed score distributions for the three models were similar in shape. The classification indices depend somewhat on the true score distribution, which is never known. When the 4PB model was fitted to the data sets, the estimated true score distributions turned out to be severely skewed. By contrast, the estimated true score distributions for the 2PB model and the transformed true score distributions of θ for the IRT model were close to the normal distribution in shape. (Note that the standard normal distribution was used as a starting point

with BILOG3 to obtain the posterior distribution of θ for the IRT model.) This does not mean that the 4PB model is worse than the other two models, because we never know the shape of the true score distribution. In fact, the 4PB model provided very good fits to the data. It should also be noted that the use of the standard normal distribution for the IRT model is arbitrary, and imposing different priors might alter the posterior distribution of θ and, in turn, the estimates of the classification accuracy indices.

The results of this study seem to suggest that model fit be examined prior to applying the estimation procedures, because the degree of the model fit is directly reflected in the estimates of the classification indices. In addition, the decision about what model to use in practice should be based, at least in part, on other considerations including suitability of the model assumptions and availability of computer programs. Higher values for estimated classification indices should not automatically dictate choice of the particular model. For example, if only model fit is considered as a criterion for choosing a model, the results of this study might support use of the IRT model. However, the randomly-parallel-form assumption of the binomial error model might be more realistic than the IRT strictly-parallel-form assumption in the sense that different forms of a test, in practice, are never strictly parallel. In other words, even though data from a single test administration are more consistent with the assumption of IRT, the measurement error associated with the binomial model assumption would be more of interest in general. For that reason, some researchers might prefer a model incorporating replications that are more flexible than strictly parallel forms (e.g., Brennan, 2000). Of course, if data are available from two administrations of a test on a representative sample of examinees, it would be preferable to use the data directly to compute the classification consistency indices (AERA, APA, & NCME, 1999, p. 35).

In order to estimate the classification accuracy indices, we need to specify true cutoff scores as well as observed cutoff scores. It is not uncommon in many testing programs that the actual cutoff scores used operationally for a test do not differ much from true score cutoffs, because the procedure for defining actual observed cutoff scores often employs a measurement model dealing with latent true scores, and all items in the pool are used (e.g., Schulz et al., 1999).

In this case, the observed cutoffs won't be much different from the true cutoffs beyond the potential differences due to roundings. In some other cases where true cutoffs do not exist specifically, the actual observed score cutoffs could be used as estimates of true score cutoffs to compute the classification accuracy indices. Or alternatively, one can find the true cutoff scores corresponding to the observed cutoffs through a mapping procedure using distributions of observed and true scores. However, it must be noted that equalizing the true and observed score cutoffs does not necessarily provide the optimal classification system, because it does not guarantee that the probability of accurate classifications is higher than any of the two error rates and that the two error rates, overall, are approximately equal.

As a final note, the cutoff scores for the examples used in this paper were expressed on the raw score metric. Since the primary score scale reported for most large scale tests are scale scores, such as percentile ranks and grade equivalents, it seems sensible to define cutoff scores on the metric of scale scores. Under the assumption that there exists a conversion table that transforms raw scores to scale scores, the raw score cutoffs corresponding to the scale score cutoffs could be found from the conversion table, and the procedures discussed in this paper could be applied to those situations.

References

ACT, Inc. (1998). *Characteristics of the Work Keys assessments*. Iowa City, IA: Author.

ACT, Inc. (1999). *Work Keys interim technical handbook*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A Guide to Criterion-Referenced Test Construction*. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*, 215-235.

Brennan, R. L. (1981). *Some statistical procedures for domain-referenced testing: A handbook for practitioners* (ACT Technical Bulletin No. 38). Iowa City, IA: ACT, Inc.

Brennan, R. L. (2000, April). *An essay on the history and future of reliability from the perspective of replications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Research Report 91-5). Iowa City, IA: ACT, Inc.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika, 43*, 317-325.

Huynh, H. (1980). Statistical inference for false positive and false negative error rates in mastery testing. *Psychometrika, 45*, 107-120.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15*, 353-368.

Huynh, H., & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17*, 351-358.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution of mental test scores. *Psychometrika, 27*, 59-72.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Lee, W., & Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1-20.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239-270.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement, 21*, 239-243.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 453-461.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG3. Item and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in Fortran* (2nd ed.). New York: Cambridge University Press.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT*. Iowa City, IA: ACT, Inc.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1997). *A study of modified-Guttman and IRT-based level scoring procedures for Work Keys Assessments* (ACT Research Report 97-7). Iowa City, IA: ACT, Inc.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*, 347-362.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-467.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore: Johns Hopkins University Press.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), *A Guide to Criterion-Referenced Test Construction*. Baltimore, MD: Johns Hopkins University Press.

Wang, T., Kolen, M. J., & Harris, D. J. (1996, April). *Conditional standard errors, reliability and decision consistency of performance levels using polytomous IRT*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics, 2*, 289-307.

TABLE 1

Theta Cutoffs and Form-Specific True and Observed Score Cutoffs

Locating Information ($K = 32$)									
		Form A		Form B		Form C		Form D	
Level	θ_h	τ_h	c_h	τ_h	c_h	τ_h	c_h	τ_h	c_h
2	-0.68	.52	16(.50)	.46	15(.47)	.48	15(.47)	.46	15(.47)
3	0.27	.66	21(.66)	.60	19(.59)	.61	19(.59)	.60	19(.59)
4	1.73	.87	27(.84)	.78	25(.78)	.79	25(.78)	.81	26(.81)
5	3.89	.99	31(.97)	.96	31(.97)	.97	31(.97)	.99	31(.97)

Applied Mathematics ($K = 30$)							
		Form X		Form Y		Form Z	
Level	θ_h	τ_h	c_h	τ_h	c_h	τ_h	c_h
2	-0.98	.41	12(.40)	.41	12(.40)	.41	12(.40)
3	-0.03	.57	17(.57)	.58	17(.57)	.57	17(.57)
4	0.73	.69	21(.70)	.70	21(.70)	.69	21(.70)
5	1.78	.86	25(.83)	.82	24(.80)	.84	25(.83)
6	2.66	.96	29(.97)	.89	28(.93)	.94	29(.97)

Note: θ_h = IRT θ cutoffs; τ_h = true score cutoffs; c_h = observed score cutoffs. The numbers in parentheses are the observed proportion-correct score cutoffs.

TABLE 2

Observed and Estimated Proportions of Examinees for Locating Information

Form	Level	Actual	IRT	4PB	2PB
A	1	.197	**.191**	.214	.226
	2	.314	.330	.282	**.310**
	3	.411	**.396**	.432	.361
	4	.073	**.075**	**.071**	.096
	5	.005	.008	.001	**.007**
B	1	.204	**.224**	.225	.253
	2	.285	.281	.253	**.282**
	3	.451	.423	**.464**	.370
	4	.060	.072	**.058**	.095
	5	.000	.001	**.000**	.001
C	1	.185	**.196**	.207	.225
	2	.290	**.284**	.251	.283
	3	.454	**.441**	.472	.391
	4	.071	.079	**.069**	.100
	5	.000	.001	**.000**	.001
D	1	.226	**.227**	.244	.248
	2	.292	**.295**	.274	.284
	3	.428	.423	**.431**	.407
	4	.052	**.051**	**.051**	.059
	5	.002	**.003**	.000	**.001**

TABLE 3

Observed and Estimated Proportions of Examinees for Applied Mathematics

Form	Level	Actual	IRT	4PB	2PB
X	1	.150	**.150**	.159	.158
	2	.278	**.282**	**.274**	.286
	3	.302	**.297**	.275	.272
	4	.194	**.195**	.216	.202
	5	.068	**.069**	.073	.078
	6	.008	**.009**	.002	.005
Y	1	.143	**.141**	**.141**	.136
	2	.251	**.263**	**.263**	.296
	3	.322	**.310**	.307	.291
	4	.194	**.191**	.199	.167
	5	.085	.090	**.088**	.099
	6	.005	**.005**	.002	.010
Z	1	.144	**.148**	.155	.152
	2	.279	.284	**.275**	.294
	3	.313	**.304**	.286	.280
	4	.198	**.196**	.222	**.200**
	5	.064	**.063**	.061	.070
	6	.002	.005	**.001**	.004

TABLE 4

Estimated Classification Consistency Indices for Locating Information

Form	Cutoff Type	IRT			4PB			2PB		
		P	κ	P_c	P	κ	P_c	P	κ	P_c
A	All	.62	.45	.31	.57	.37	.32	.55	.36	.29
	$m = 1$	.89	.65	.69	.87	.61	.66	.84	.54	.65
	$m = 2$	.79	.58	.50	.78	.55	.50	.78	.56	.50
	$m = 3$	.93	.56	.85	.89	.16	.87	.90	.45	.82
	$m = 4$	.99	.34	.99	1.0	.00	1.0	.99	.20	.99
B	All	.60	.41	.31	.55	.33	.33	.52	.32	.29
	$m = 1$	.86	.61	.65	.86	.59	.65	.81	.49	.62
	$m = 2$	.78	.57	.50	.73	.47	.50	.76	.51	.50
	$m = 3$	.92	.43	.87	.90	.08	.89	.90	.40	.83
	$m = 4$	1.0	.15	1.0	1.0	.00	1.0	1.0	.09	1.0
C	All	.60	.41	.32	.54	.31	.33	.51	.30	.29
	$m = 1$	.87	.60	.69	.86	.57	.67	.81	.46	.65
	$m = 2$	.78	.55	.50	.74	.48	.50	.75	.49	.50
	$m = 3$	.92	.46	.85	.88	.10	.87	.89	.39	.82
	$m = 4$	1.0	.19	1.0	1.0	.00	1.0	1.0	.08	1.0
D	All	.61	.43	.32	.55	.33	.32	.54	.33	.31
	$m = 1$	.85	.58	.65	.82	.51	.63	.81	.48	.63
	$m = 2$	.78	.56	.50	.75	.50	.50	.75	.50	.50
	$m = 3$	.95	.53	.90	.92	.22	.90	.93	.35	.89
	$m = 4$	1.0	.32	.99	1.0	.00	1.0	1.0	.08	1.0

TABLE 5

Estimated Classification Accuracy Indices for Locating Information

Form	Cutoff Type	IRT			4PB			2PB		
		γ	P^+	P^-	γ	P^+	P^-	γ	P^+	P^-
A	All	.70	.22	.08	.67	.23	.10	.63	.27	.10
	$m = 1$	.91	.07	.02	.90	.06	.04	.88	.08	.05
	$m = 2$	.85	.10	.06	.84	.10	.07	.84	.11	.05
	$m = 3$	.95	.05	.00	.93	.07	.00	.92	.08	.00
	$m = 4$	.99	.01	.00	1.0	.00	.00	.99	.01	.00
B	All	.69	.20	.10	.66	.21	.14	.61	.26	.13
	$m = 1$	.90	.05	.04	.90	.04	.06	.86	.06	.08
	$m = 2$	.84	.11	.05	.80	.11	.09	.81	.14	.05
	$m = 3$	.95	.04	.01	.94	.06	.00	.93	.06	.01
	$m = 4$	1.0	.00	.00	1.0	.00	.00	1.0	.00	.00
C	All	.68	.24	.08	.64	.24	.12	.59	.29	.11
	$m = 1$	.90	.07	.03	.90	.06	.05	.86	.08	.06
	$m = 2$	.83	.13	.05	.80	.13	.07	.80	.15	.05
	$m = 3$	.95	.05	.01	.93	.07	.00	.92	.07	.01
	$m = 4$	1.0	.00	.00	1.0	.00	.00	1.0	.00	.00
D	All	.71	.18	.12	.65	.22	.14	.63	.22	.14
	$m = 1$	.89	.06	.05	.87	.06	.07	.86	.06	.08
	$m = 2$	.84	.10	.06	.82	.12	.07	.81	.12	.06
	$m = 3$	.97	.02	.01	.95	.05	.00	.95	.05	.00
	$m = 4$	1.0	.00	.00	1.0	.00	.00	1.0	.00	.00

TABLE 6

Estimated Classification Consistency Indices for Applied Mathematics

Form	Cutoff Type	IRT			4PB			2PB		
		P	κ	P_c	P	κ	P_c	P	κ	P_c
X	All	.55	.41	.23	.46	.30	.23	.46	.30	.23
	$m = 1$	.91	.64	.75	.88	.54	.73	.87	.51	.73
	$m = 2$	.82	.63	.51	.79	.58	.51	.79	.57	.51
	$m = 3$	.85	.62	.60	.80	.53	.59	.81	.54	.59
	$m = 4$	.94	.60	.86	.91	.35	.86	.91	.44	.85
	$m = 5$	.99	.44	.98	1.0	.04	1.0	.99	.19	.99
Y	All	.52	.38	.23	.43	.26	.23	.43	.27	.23
	$m = 1$	.91	.63	.76	.89	.55	.76	.87	.47	.76
	$m = 2$	.83	.64	.52	.79	.56	.52	.77	.53	.51
	$m = 3$	.83	.58	.59	.77	.43	.59	.80	.51	.60
	$m = 4$	.91	.49	.83	.87	.22	.84	.89	.43	.81
	$m = 5$	.99	.22	.99	1.0	.01	1.0	.98	.22	.98
Z	All	.54	.39	.24	.46	.29	.23	.46	.29	.23
	$m = 1$	.91	.63	.75	.88	.54	.74	.87	.50	.74
	$m = 2$	.81	.62	.51	.79	.57	.51	.78	.55	.51
	$m = 3$	.85	.61	.61	.79	.49	.59	.81	.53	.60
	$m = 4$	.94	.55	.87	.91	.26	.88	.92	.41	.86
	$m = 5$	.99	.33	.99	1.0	.01	1.0	.99	.16	.99

TABLE 7

Estimated Classification Accuracy Indices for Applied Mathematics

Form	Cutoff Type	IRT			4PB			2PB		
		γ	P^+	P^-	γ	P^+	P^-	γ	P^+	P^-
X	All	.64	.26	.10	.57	.28	.15	.57	.28	.15
	$m = 1$	.94	.03	.03	.91	.05	.04	.91	.05	.05
	$m = 2$	.86	.09	.04	.84	.09	.06	.84	.10	.06
	$m = 3$	.89	.09	.02	.86	.09	.05	.87	.09	.05
	$m = 4$	.95	.05	.00	.93	.07	.00	.93	.07	.00
	$m = 5$	.99	.01	.00	1.0	.00	.00	.99	.01	.00
Y	All	.61	.31	.09	.53	.32	.15	.53	.33	.14
	$m = 1$	.93	.04	.03	.92	.04	.04	.91	.05	.04
	$m = 2$	.86	.11	.03	.84	.12	.06	.82	.12	.05
	$m = 3$	.87	.11	.02	.83	.11	.06	.86	.10	.04
	$m = 4$	.93	.07	.00	.91	.09	.00	.92	.08	.01
	$m = 5$	1.0	.00	.00	1.0	.00	.00	.99	.01	.00
Z	All	.64	.25	.11	.57	.27	.16	.57	.27	.16
	$m = 1$	.93	.03	.03	.91	.05	.04	.90	.05	.05
	$m = 2$	.86	.10	.04	.84	.09	.07	.84	.10	.06
	$m = 3$	.89	.09	.03	.85	.09	.06	.86	.09	.05
	$m = 4$	.96	.04	.00	.94	.06	.00	.94	.05	.00
	$m = 5$	1.0	.00	.00	1.0	.00	.00	1.0	.00	.00

FIGURE 1. **Classification Consistency and Accuracy Indices With $H = 4$**

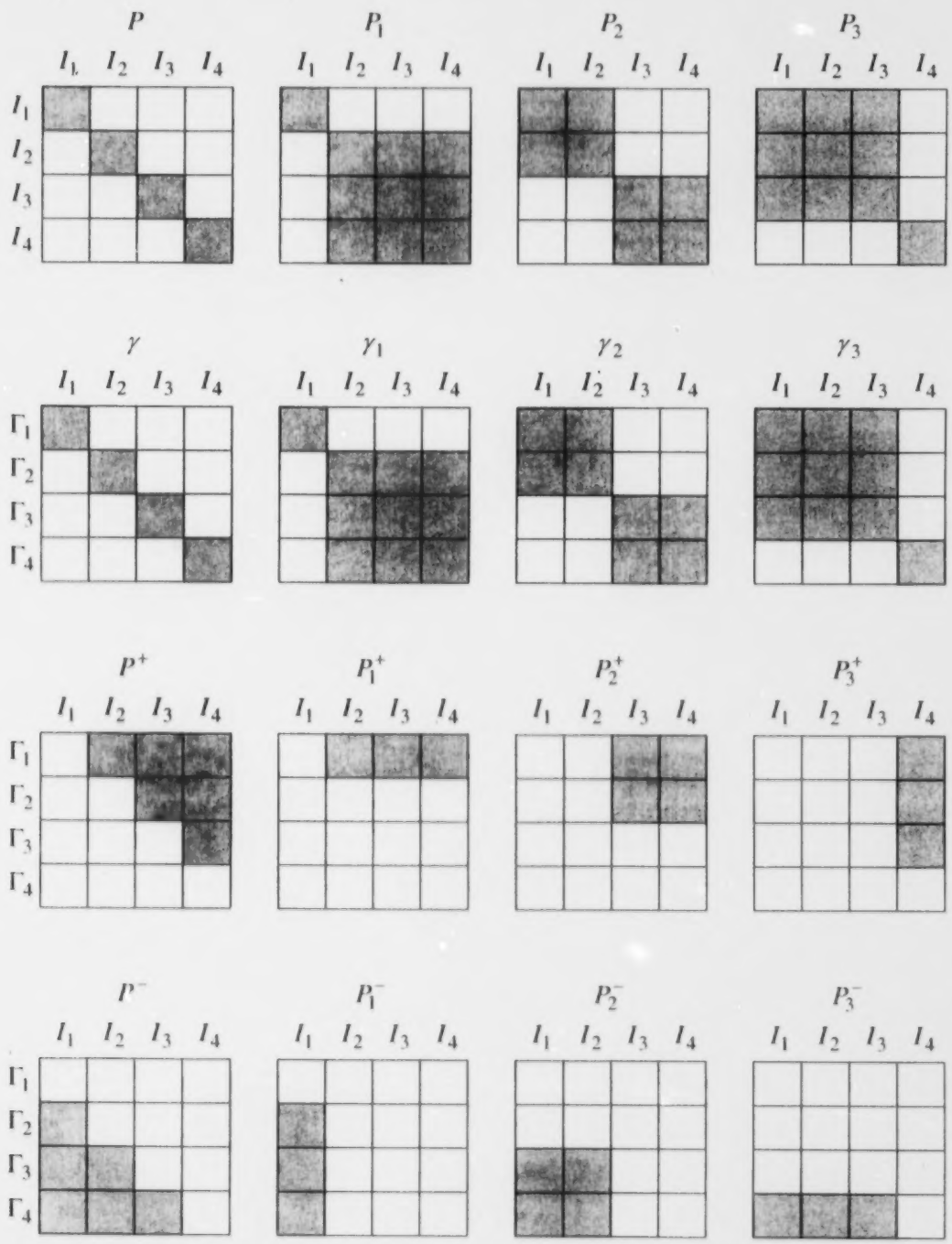

FIGURE 2. **Model Fit for Locating Information**

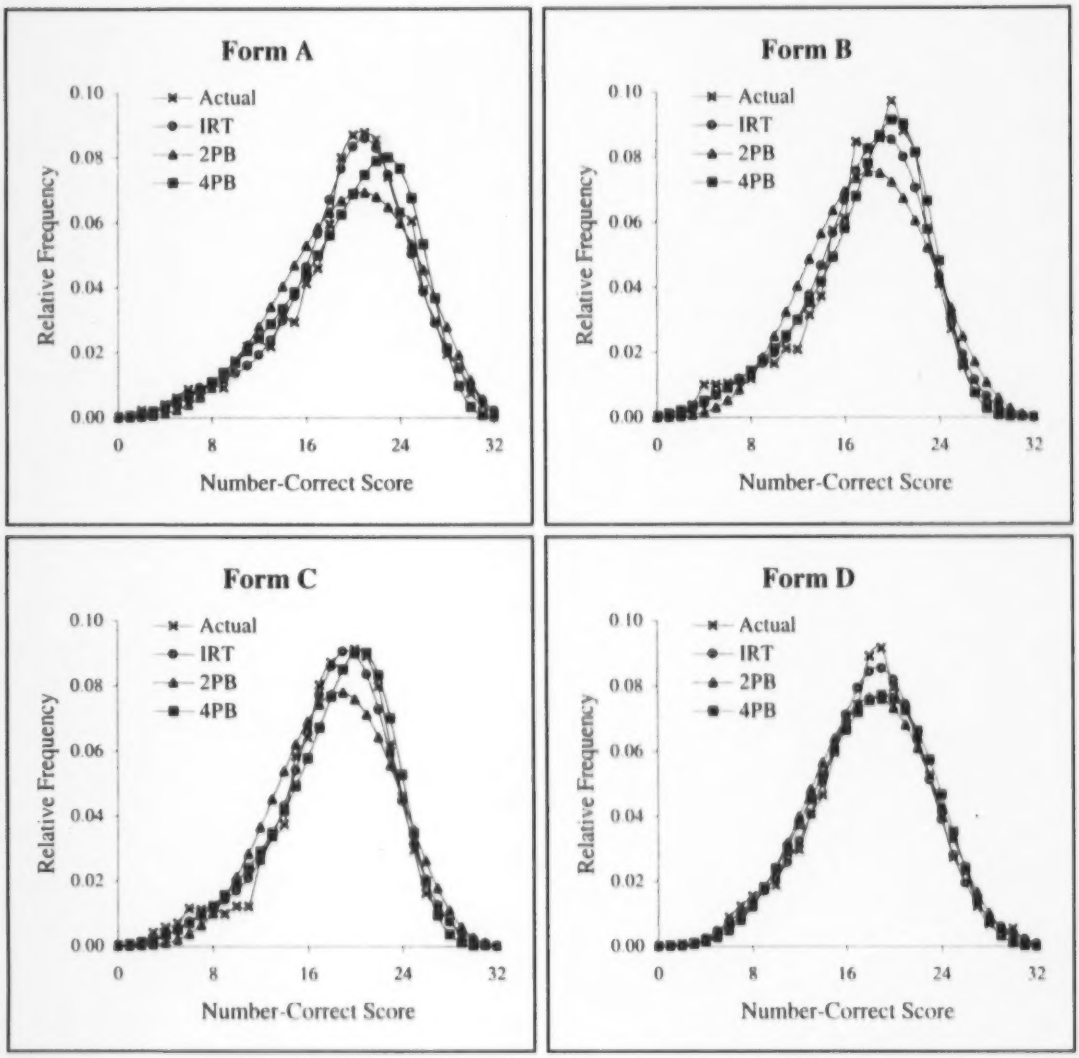

FIGURE 3. **Model Fit for Applied Mathematics**

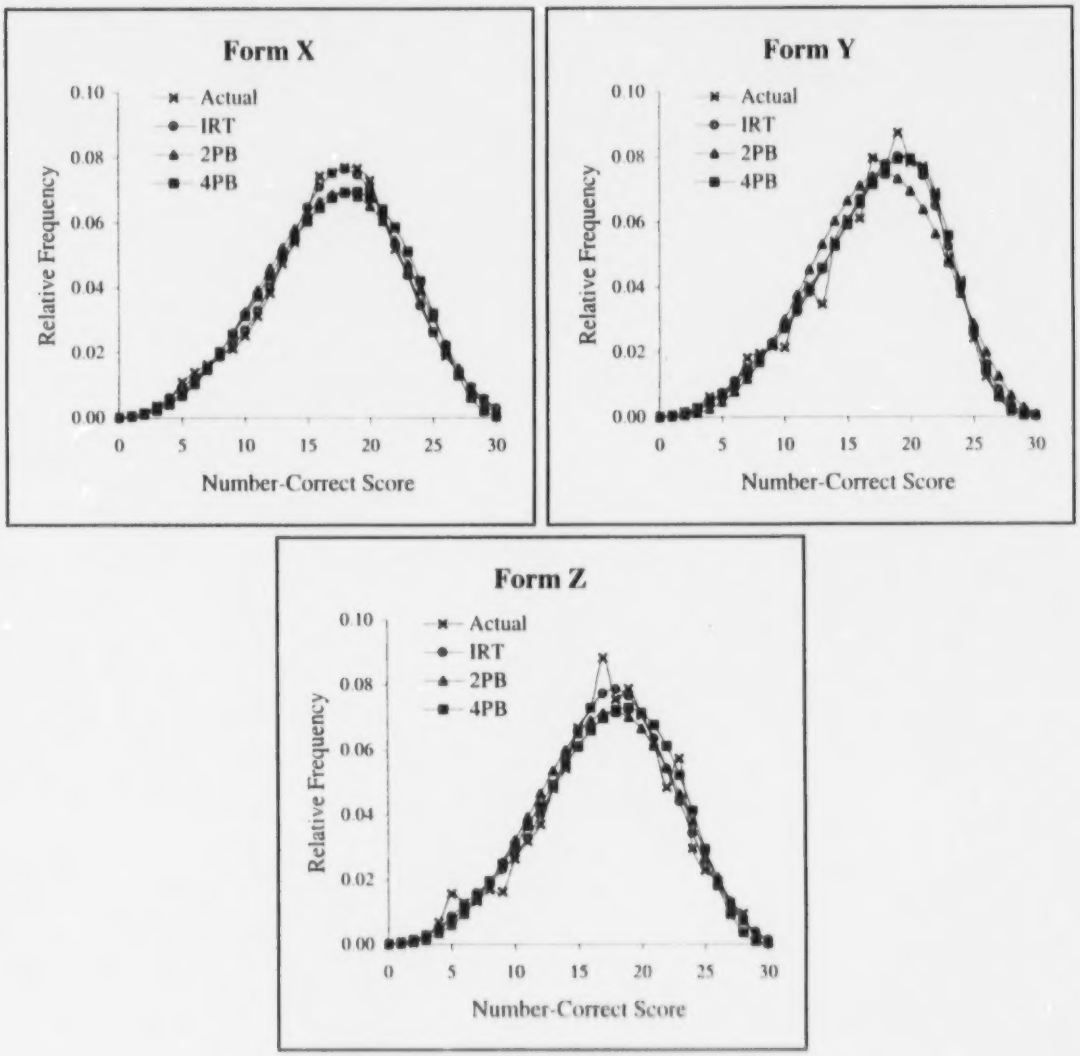

FIGURE 4. **Estimated True Score Distributions for Locating Information**

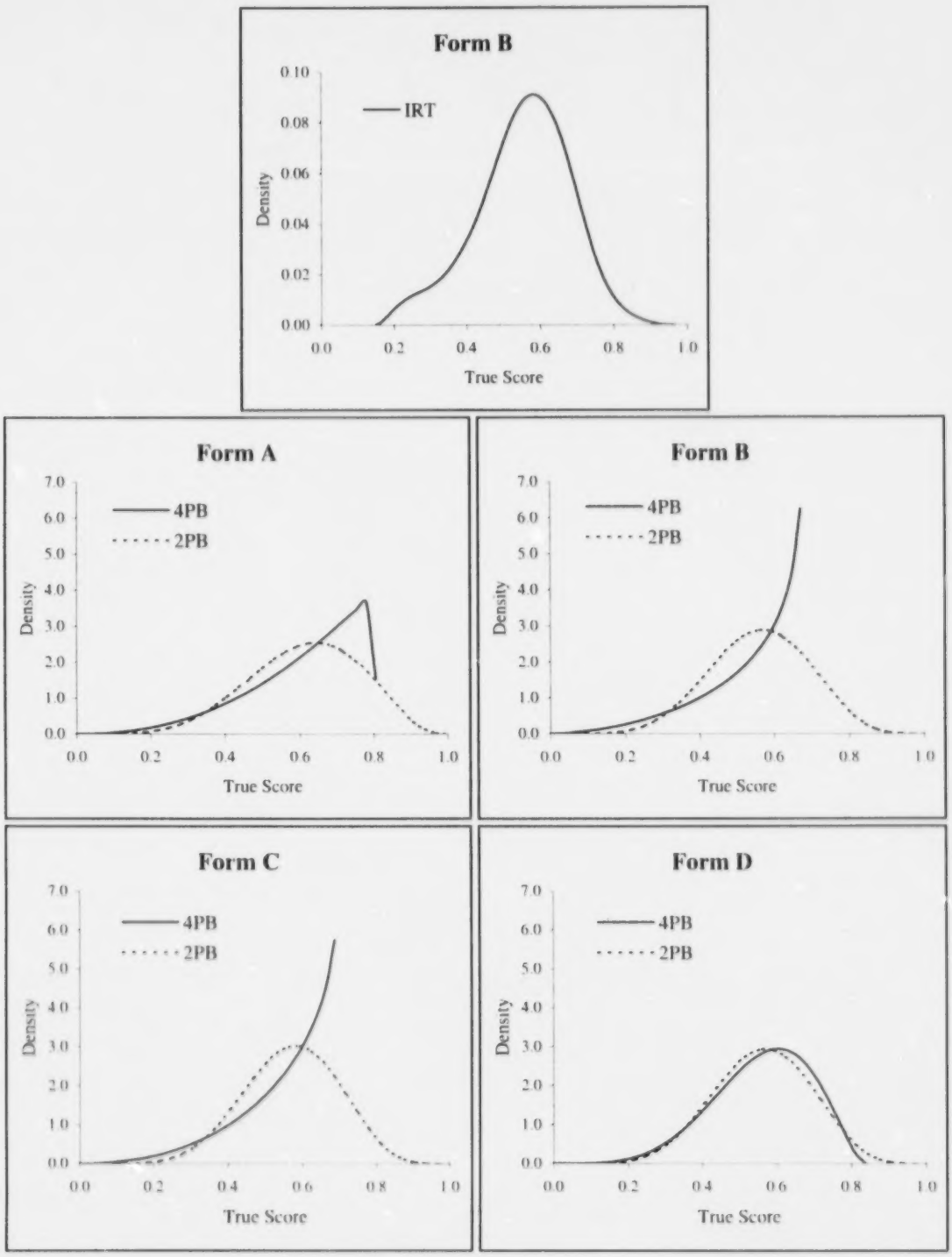

FIGURE 5. **Estimated True Score Distributions for Applied Mathematics**

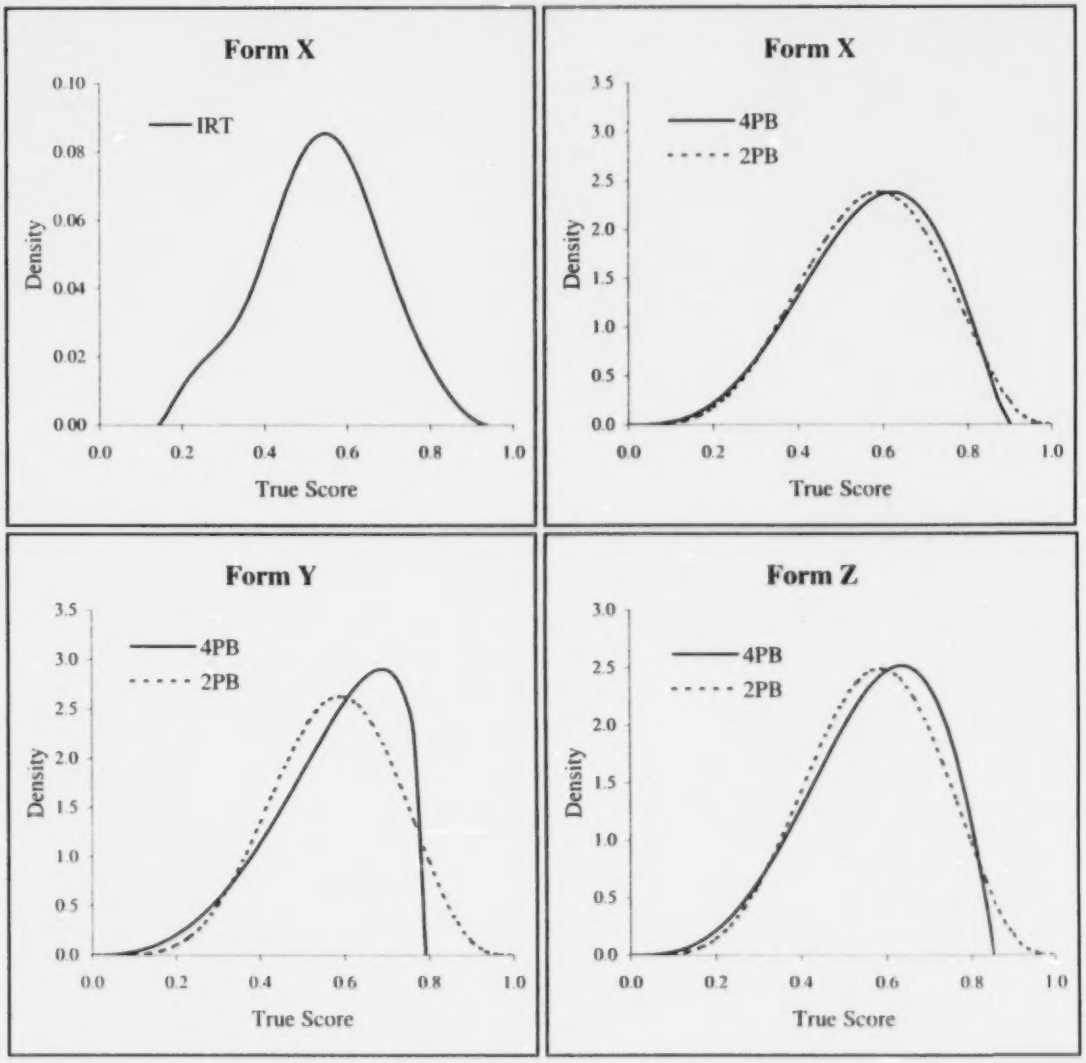

FIGURE 6. **Estimated Conditional Probabilities of Inconsistent Classifications for Locating Information**

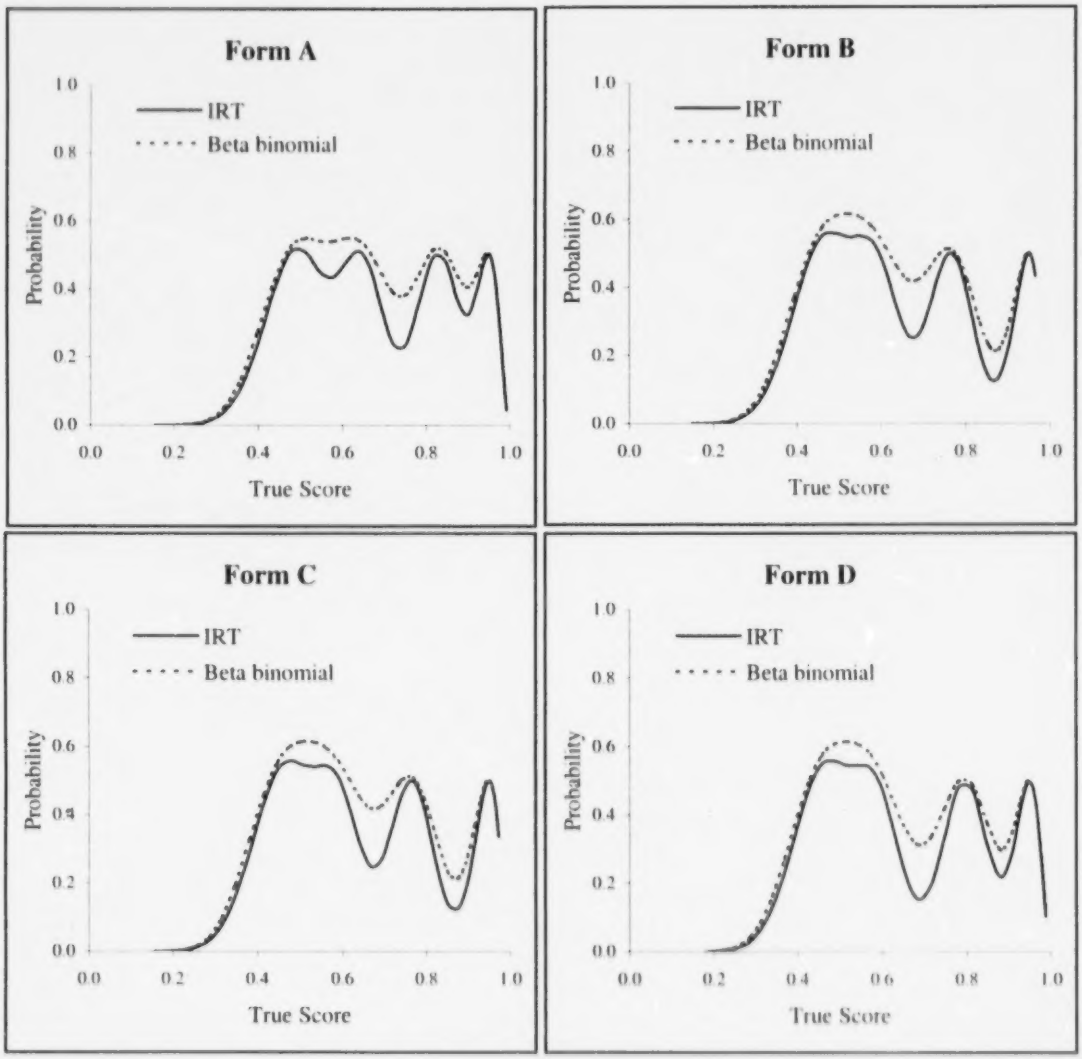

FIGURE 7. **Estimated Conditional Probabilities of Inconsistent Classifications for Applied Mathematics**

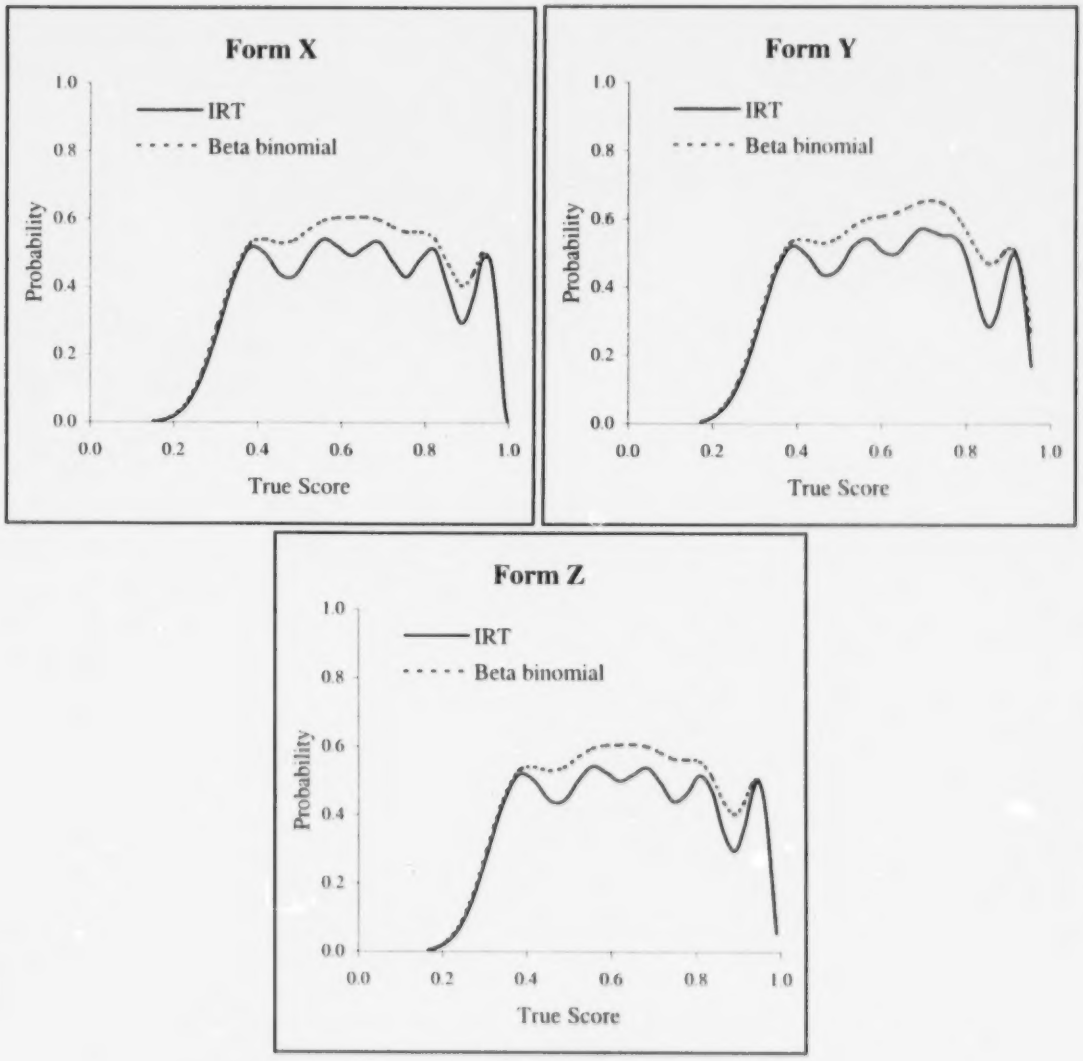

FIGURE 8. **Estimated Conditional Probabilities of Observed Categories Using IRT Model for Locating Information**

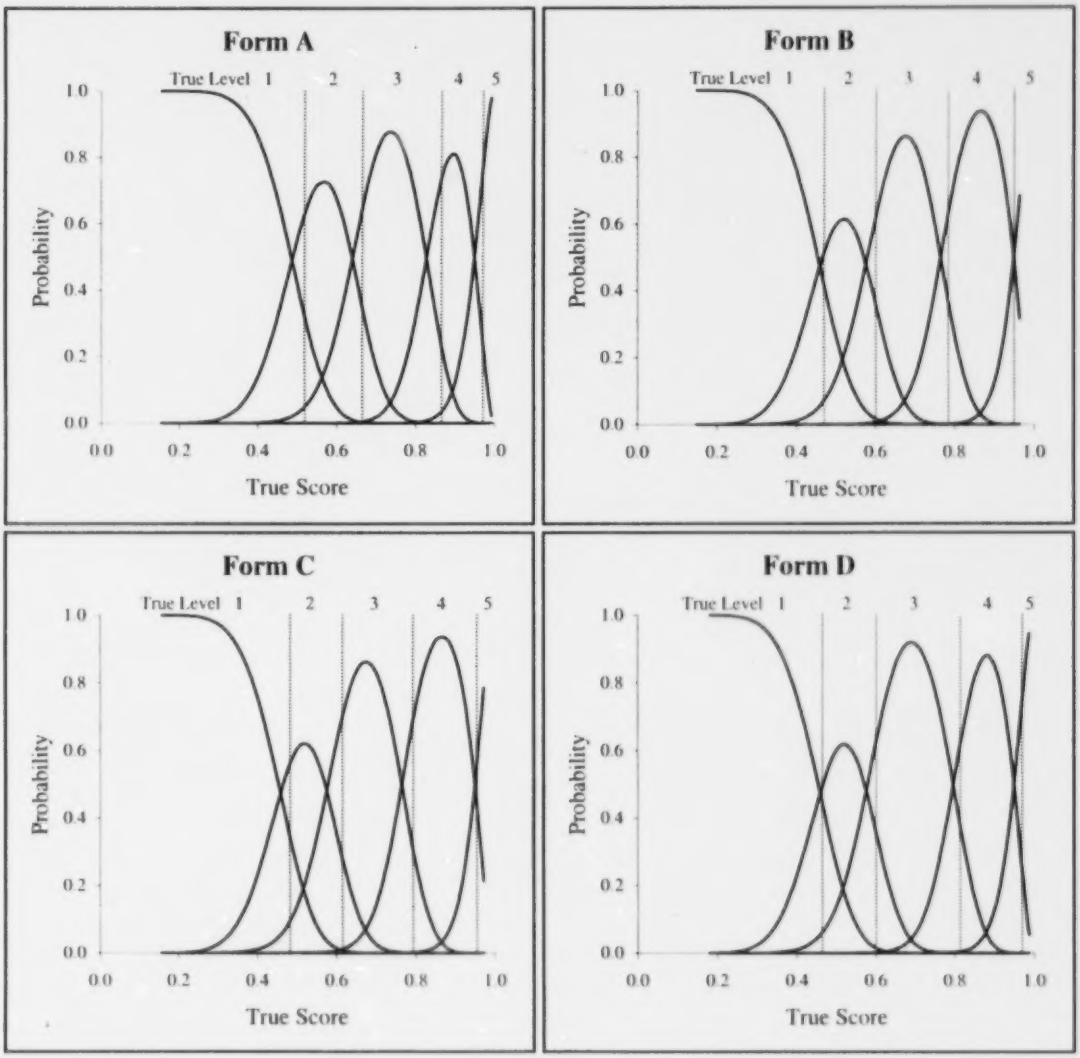

FIGURE 9. **Estimated Conditional Probabilities of Observed Categories Using Beta Binomial Model for Locating Information**

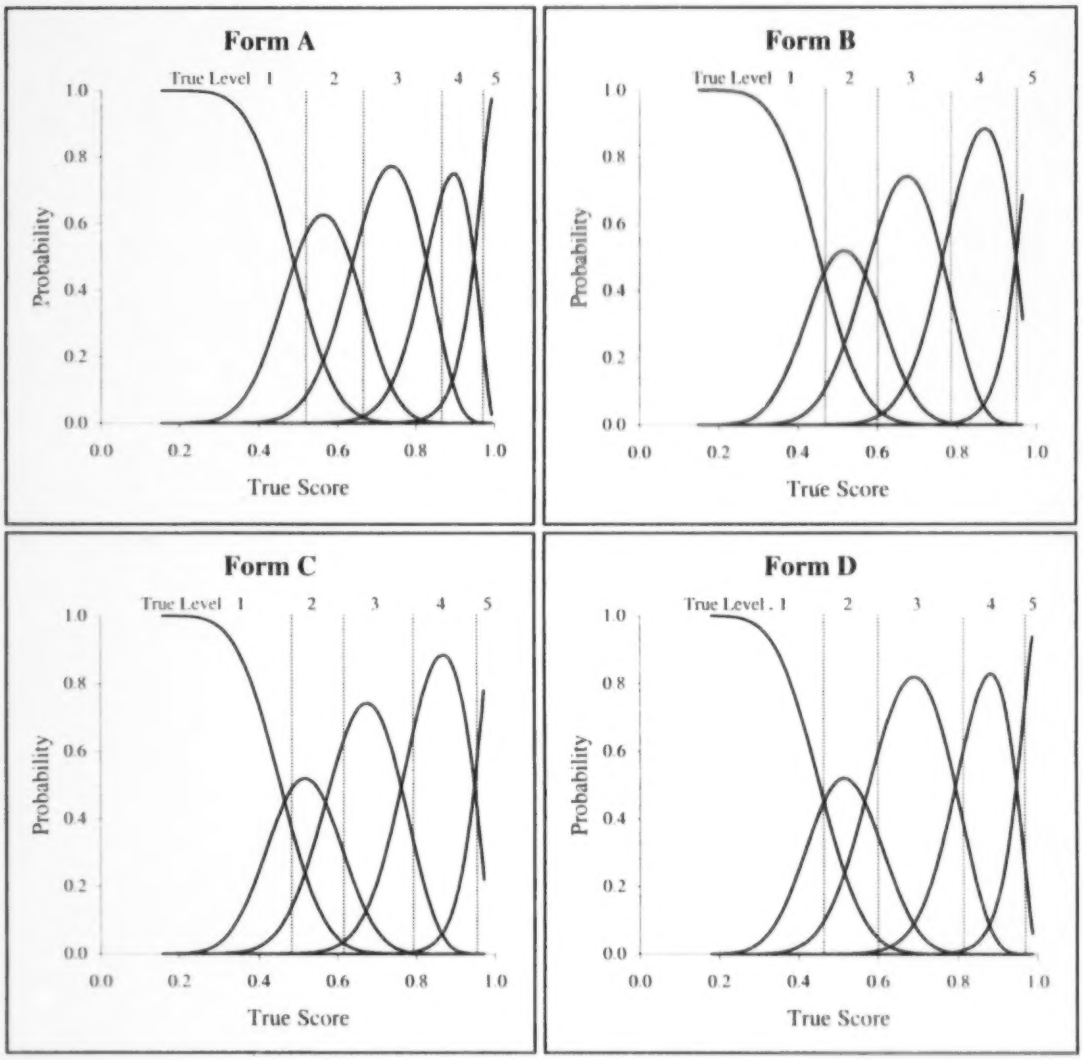

FIGURE 10. **Estimated Conditional Probabilities of Observed Categories Using IRT Model for Applied Mathematics**

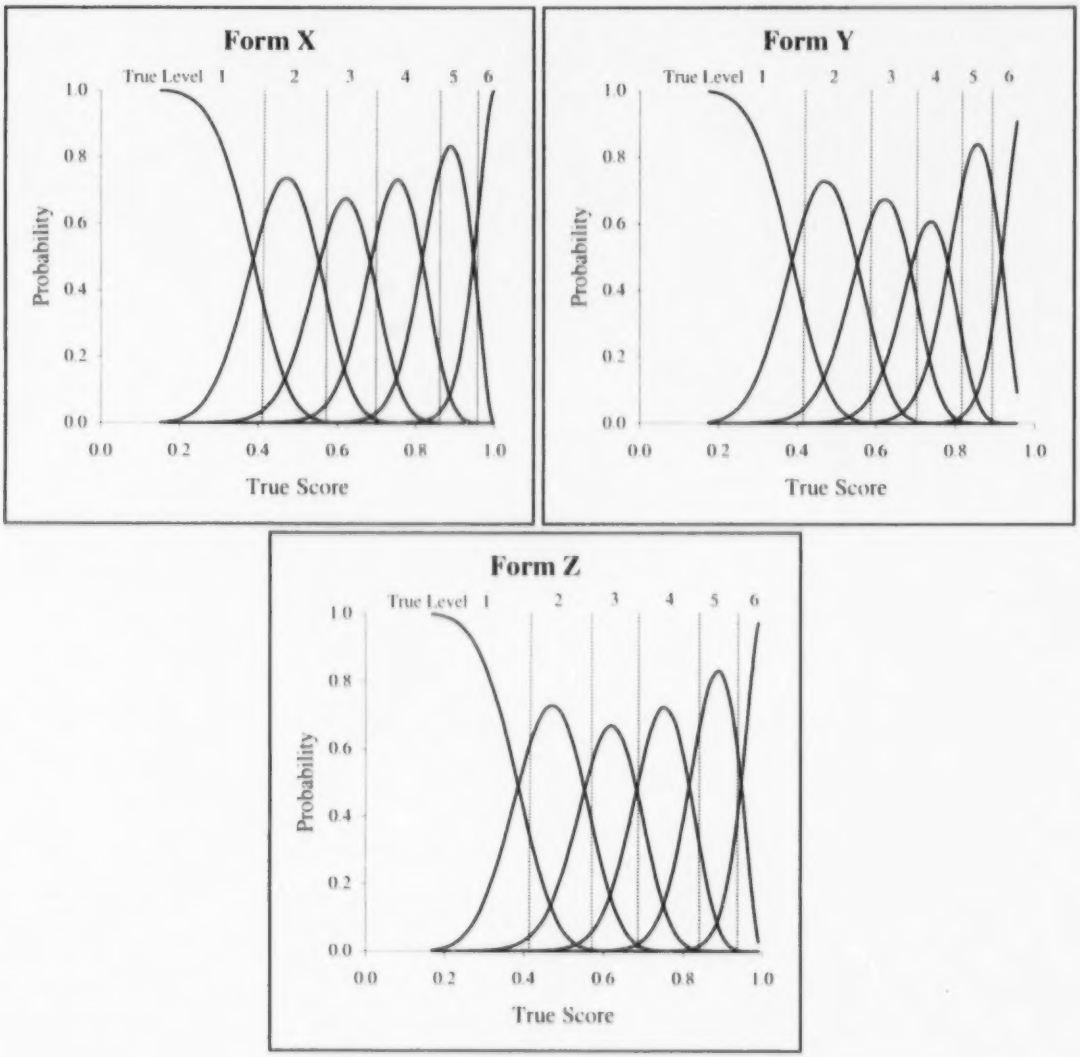

FIGURE 11. **Estimated Conditional Probabilities of Observed Categories Using Beta Binomial Model for Applied Mathematics**

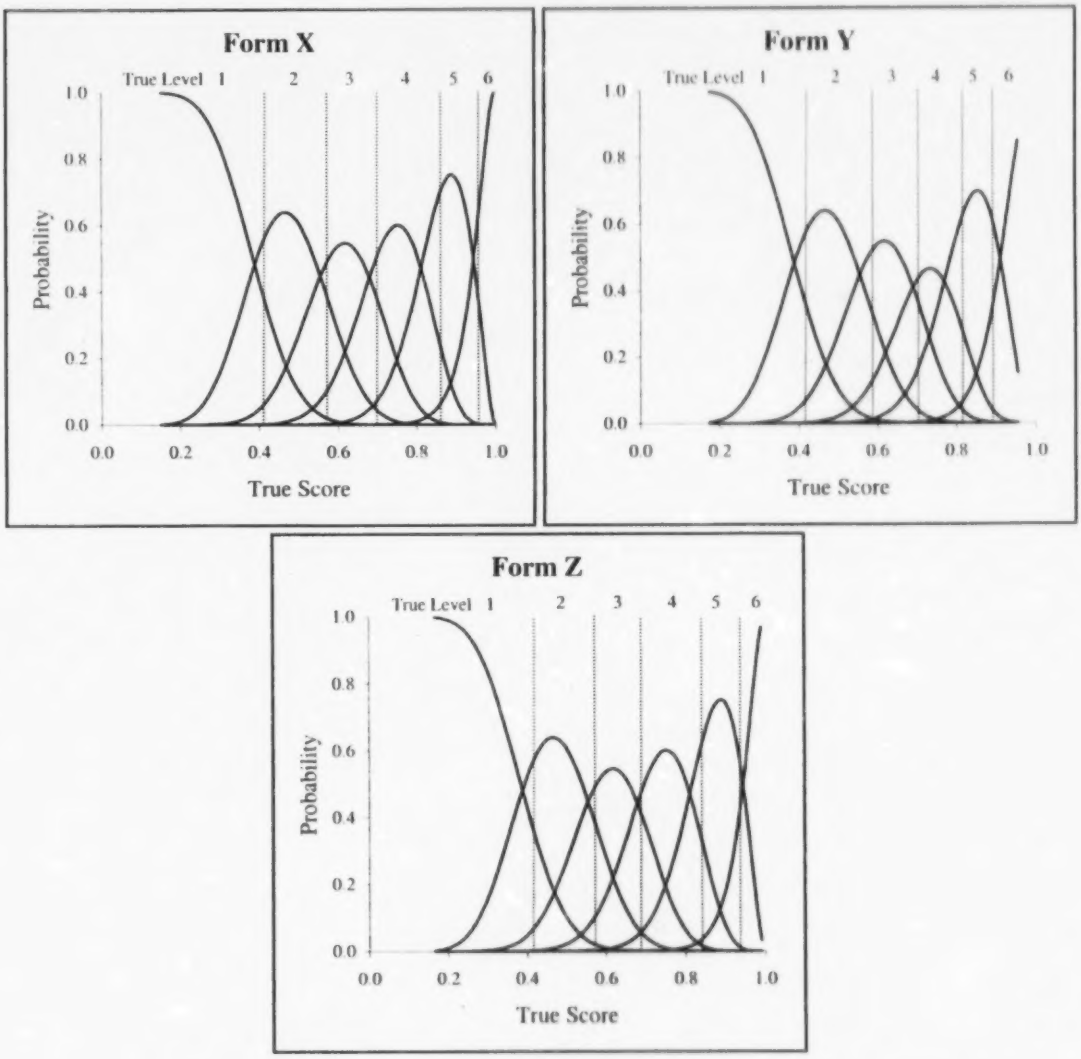

FIGURE 12. **Estimated Conditional Probabilities of Accurate, Lower Than True Level, and Higher Than True Level Classifications Using IRT Model for Locating Information**

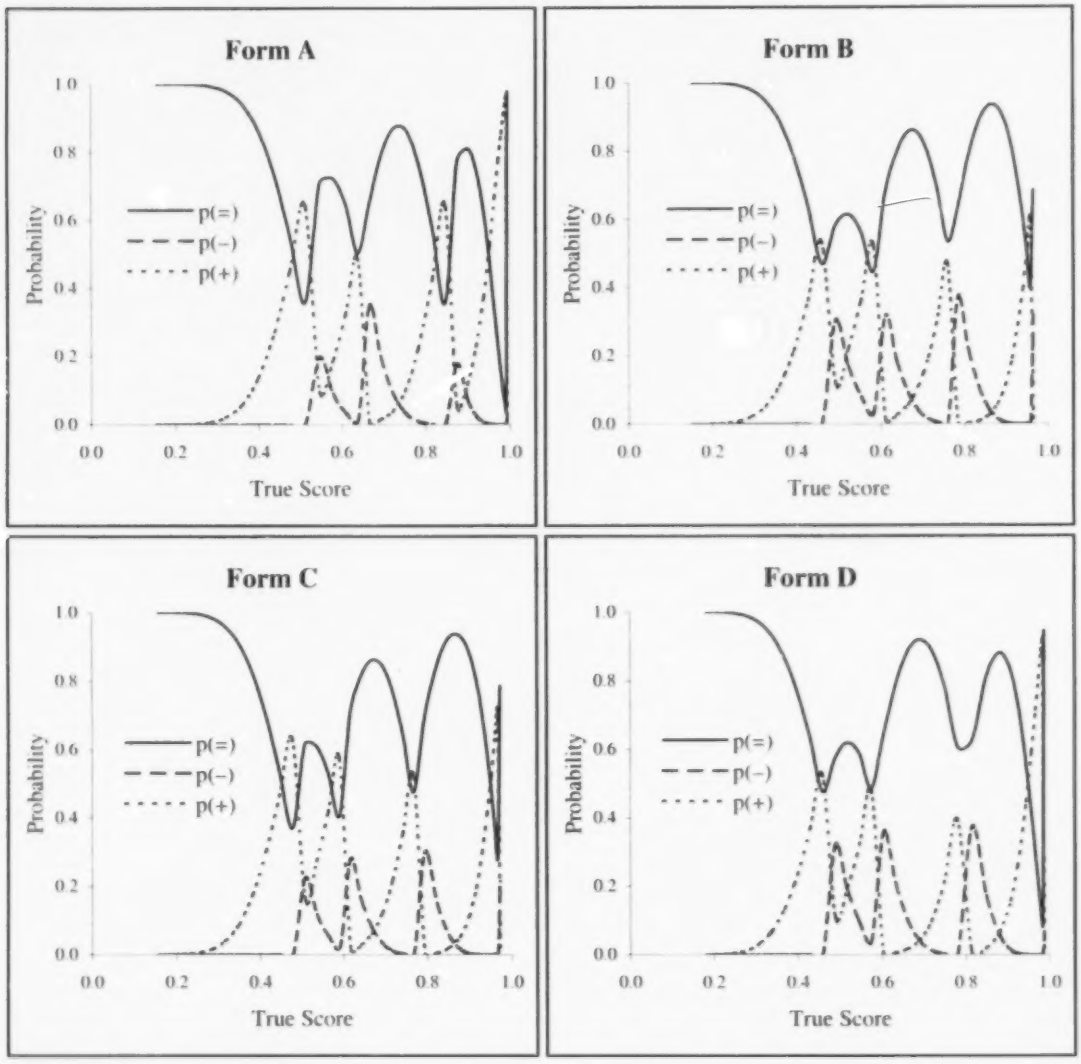

FIGURE 13. **Estimated Conditional Probabilities of Accurate, Lower Than True Level, and Higher Than True Level Classifications Using Beta Binomial Model for Locating Information**

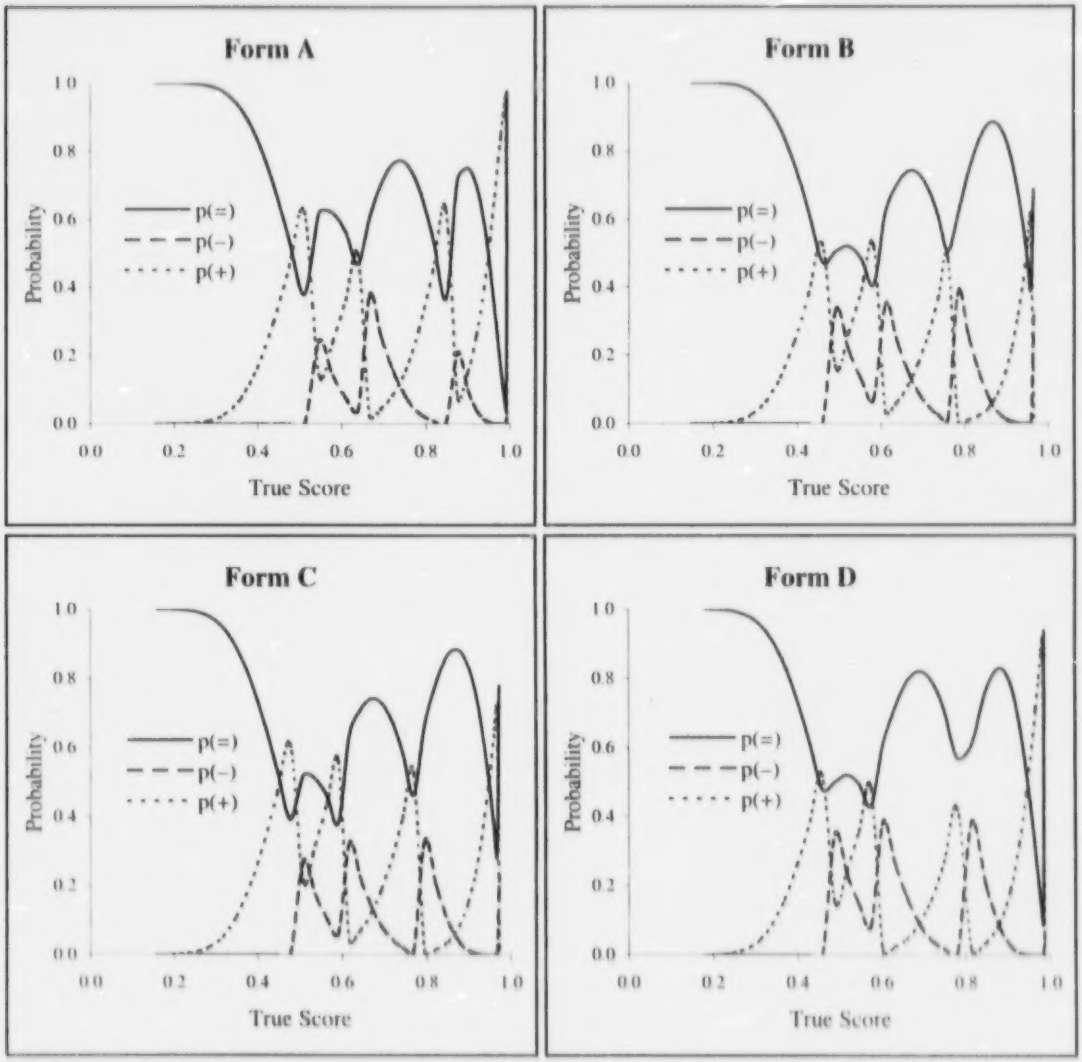

FIGURE 14. **Estimated Conditional Probabilities of Accurate, Lower Than True Level, and Higher Than True Level Classifications Using IRT Model for Applied Mathematics**

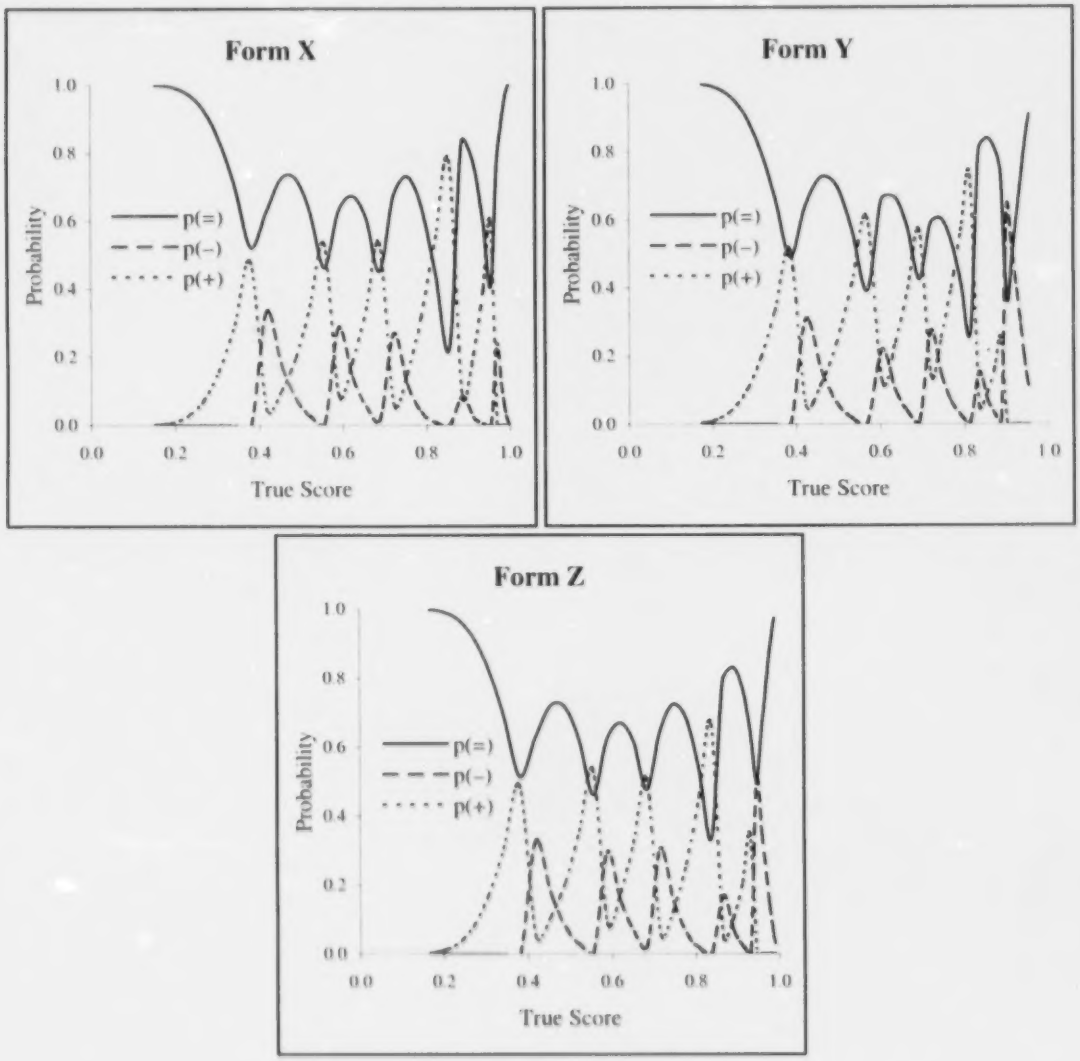

FIGURE 15. **Estimated Conditional Probabilities of Accurate, Lower Than True Level, and Higher Than True Level Classifications Using Beta Binomial Model for Applied Mathematics**

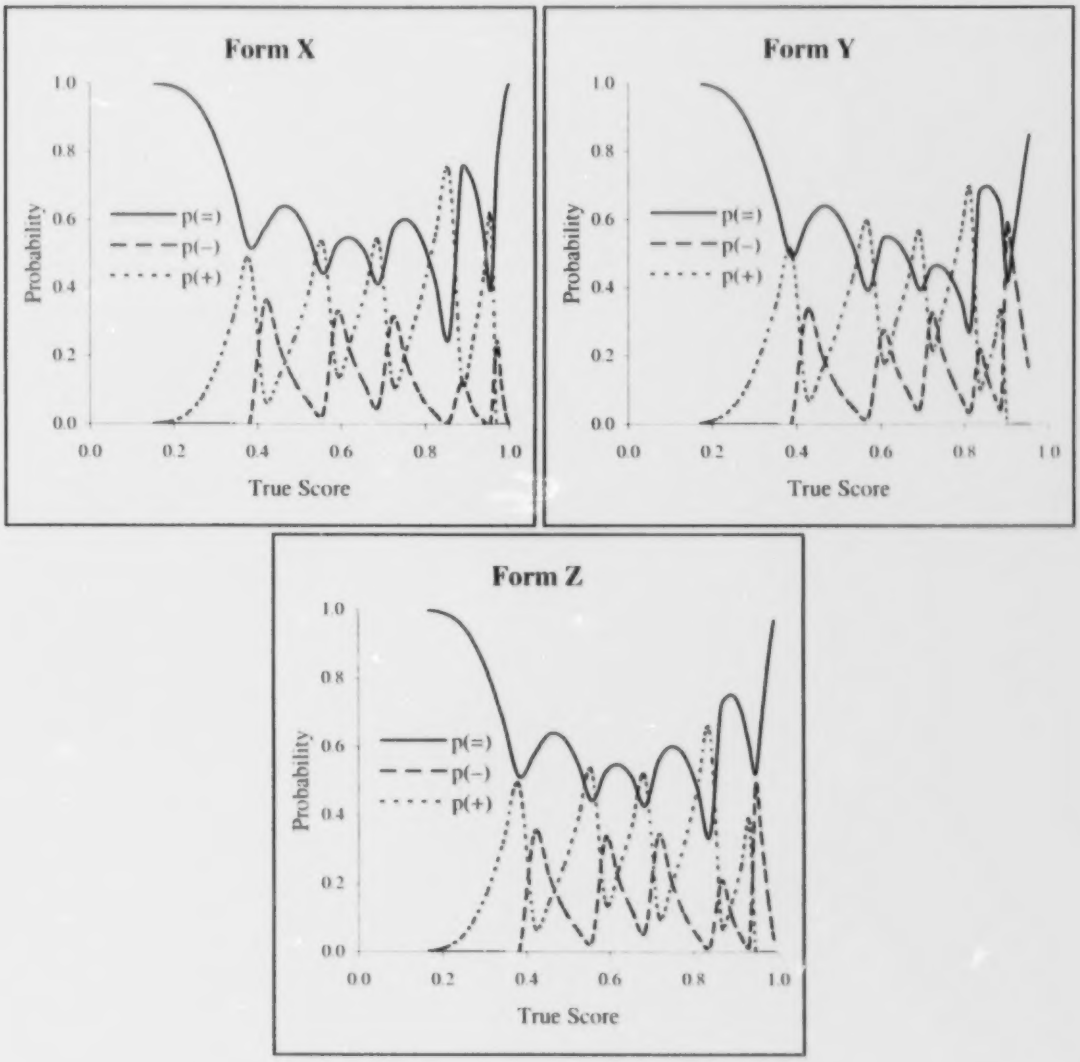

FIGURE 16. **Selected Conditional Probabilities of Observed Scores with True Score Cutoffs Using IRT Model for Form Y, Applied Mathematics**